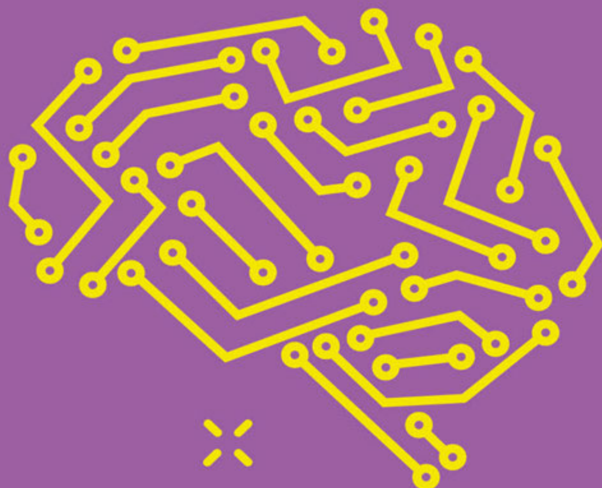
Inteligencia Artificial

MARGARET A. BODEN



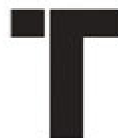
TURNER NOEMA



Inteligencia artificial

MARGARET A. BODEN

TRADUCCIÓN DE INMACULADA PÉREZ PARRA



COLECCIÓN NOEMA

Título:

Inteligencia artificial

Edición original en inglés: AI. Its Nature and Future

Margaret A. Boden, 2016

De esta edición:

Diego de León, 30

28006 Madrid

www.turnerlibros.com

Primera edición: octubre de 2017

ISBN: 978-84-16714-22-3

Diseño de la colección:

Enric Satué

Ilustración de cubierta:

Diseño TURNER

Depósito Legal: M-28740-2017
Impreso en España

La editorial agradece todos los comentarios y observaciones:

turner@turnerlibros.com

Para Byron, Oscar, Lukas y Alina.

ÍNDICE


MEMES_DEL_SARGENTO.PIMIENTA

I

¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

La inteligencia artificial (IA) tiene por objeto que los ordenadores hagan la misma clase de cosas que puede hacer la mente.

Algunas (como razonar) se suelen describir como “inteligentes”. Otras (como la visión), no. Pero todas entrañan competencias psicológicas (como la percepción, la asociación, la predicción, la planificación, el control motor) que permiten a los seres humanos y demás animales alcanzar sus objetivos.

La inteligencia no es una dimensión única, sino un espacio profusamente estructurado de capacidades diversas para procesar la información. Del mismo modo, la IA utiliza muchas técnicas diferentes para resolver una gran variedad de tareas.

Y está en todas partes.

Encontramos aplicaciones prácticas de la IA en el hogar, en los coches (y en los vehículos sin conductor), en las oficinas, los bancos, los hospitales, el cielo… y en internet, incluido el Internet de las Cosas (que conecta los sensores físicos cada vez más numerosos de nuestros aparatos, ropa y entorno). Algunas se salen de nuestro planeta, como los robots enviados a la Luna y a Marte o los satélites que orbitan en el espacio. Las animaciones de Hollywood, los videojuegos y los juegos de ordenador, los sistemas de navegación por satélite y el motor de búsqueda de Google están basados en técnicas de IA, igual que los sistemas que usan los financieros para predecir los movimientos del mercado de valores y los gobiernos nacionales como guía para tomar decisiones políticas sobre salud y transporte, igual que las aplicaciones de los teléfonos

móviles. Añadamos los avatares de la realidad virtual y los modelos de la emoción experimentales creados para los robots de “compañía”. Hasta las galerías de arte utilizan la IA en sus páginas web y también en las exposiciones de arte digital. Desgraciadamente, hay drones militares recorriendo los campos de batalla, pero, por suerte, también hay robots dragaminas.

La IA tiene dos objetivos principales. Uno es tecnológico: usar los ordenadores para hacer cosas útiles (a veces empleando métodos muy distintos a los de la mente). El otro es científico: usar conceptos y modelos de IA que ayuden a resolver cuestiones sobre los seres humanos y demás seres vivos. La mayoría de los especialistas en IA se concentra en un solo objetivo, aunque algunos contemplan ambos.

Además de proporcionar infinidad de chismes tecnológicos, la IA ha influido profundamente en las biociencias. Un modelo informático de una teoría científica es prueba de su claridad y coherencia y una demostración convincente de sus implicaciones (por lo general desconocidas). Que la teoría sea verdad es otro asunto, y dependerá de las pruebas obtenidas por la ciencia en cuestión, pero el modelo puede resultar esclarecedor, incluso si se demuestra que la teoría es falsa.

Concretamente, la IA ha hecho posible que psicólogos y neurocientíficos desarrollen influyentes teorías sobre la entidad mente-cerebro, incluyendo modelos de cómo funciona el cerebro físico y –pregunta distinta pero igualmente importante– qué es lo que hace el cerebro: a qué cuestiones computacionales (psicológicas) responde y qué clases de procesamiento de la información le permiten hacerlo. Quedan muchas preguntas sin responder, ya que la misma IA nos ha enseñado que la mente es mucho más rica de lo que los psicólogos se habían imaginado en un principio.

Los biólogos, también, han utilizado la IA en forma de “vida artificial” (A-Life) para desarrollar modelos computacionales de

diversos aspectos de organismos vivos que les ayudan a explicar varios tipos de comportamiento animal, el desarrollo de la forma corporal, la evolución biológica y la naturaleza de la vida misma.

Además de repercutir en las biociencias, la IA ha influido en la filosofía. Muchos filósofos actuales basan sus juicios sobre la mente en conceptos de IA; los utilizan para abordar, por ejemplo, el muy mentado problema mente-cuerpo, el enigma del libre albedrío y los muchos misterios de la conciencia. No obstante, estas ideas filosóficas son extremadamente controvertidas y existen profundas discrepancias sobre si algún sistema de IA podría poseer auténtica inteligencia, creatividad o vida.

Por último, aunque no menos importante, la IA ha puesto en entredicho nuestro concepto de la humanidad y su futuro. Algunos incluso dudan si de hecho tendremos futuro, porque prevén que la IA superará a la inteligencia humana en todos los ámbitos. Aunque algunos pensadores ven esto con agrado, la mayoría lo teme: ¿qué lugar quedará, se preguntan, para la dignidad y la responsabilidad humanas?

Todas estas cuestiones se estudiarán en los capítulos siguientes.

MÁQUINAS VIRTUALES

“Pensar en IA –podría decirse–, es pensar en ordenadores”. Bueno, sí y no. Los ordenadores, como tales, no son la cuestión; lo que hacen es lo que importa. Dicho de otro modo: aunque la IA precisa de máquinas físicas (por ejemplo, ordenadores), sería más acertado considerar que utiliza lo que los especialistas en sistemas llaman máquinas virtuales.

Una máquina virtual no es la representación de una máquina en la realidad virtual, ni se parece a un motor de coche simulado para estudiar mecánica; es más bien el sistema de procesamiento de la información que el programador concibe cuando escribe un programa y el que tiene en mente la gente al usarlo.

Como analogía, pensemos en una orquesta. Los instrumentos tienen que funcionar. Madera, metal, piel y tripas de gato deben seguir las leyes de la música para que esta suene como debiera. Pero los que asisten al concierto no se fijan en eso, están más interesados en la música. Tampoco les preocupan las notas individuales, y menos todavía las vibraciones en el aire que provocan el sonido. Están escuchando las “formas” musicales que componen las notas: melodías y armonías, temas y variaciones, ligaduras y síncopas.

En lo que respecta a la IA, la situación es similar. Por ejemplo: un procesador de textos es algo que, para el diseñador que lo concibe, y para el usuario que lo utiliza, trata directamente con palabras y párrafos. Pero, por lo general, el programa en sí mismo no contiene ninguna de esas dos cosas. (Algunos sí, como por ejemplo los avisos de copyright que el usuario puede insertar fácilmente). Y una red neuronal (véase el capítulo IV) se considera una manera de procesar información en paralelo, a pesar de que se suele aplicar (de forma secuencial) en un computador de Von Neumann.

Eso no significa que una máquina virtual sea una ficción útil, un mero producto de nuestra imaginación. Las máquinas virtuales son realidades concretas. Pueden llevar a cabo tareas, tanto dentro del sistema como en el mundo exterior (si están conectadas a dispositivos físicos como cámaras o manos robóticas). Cuando los que se dedican a la IA indagan qué no funciona en un programa que hace algo inesperado, rara vez contemplan fallos en el hardware. Por lo general, se interesan por las reacciones e interacciones causales en la maquinaria virtual, el software.

Los lenguajes de programación también son máquinas virtuales (cuyas instrucciones tienen que ser traducidas al código de la máquina para que las pueda llevar a cabo). Algunos se definen en función de lenguajes de programación de bajo nivel, por lo que requieren una traducción a varios niveles. Los lenguajes de programación son necesarios porque muy pocas personas pueden

procesar la información con la configuración de bits que utiliza el código máquina y porque nadie puede pensar en procesos complejos a un nivel tan detallado.

Los lenguajes de programación no son el único caso de máquinas virtuales. Las máquinas virtuales suelen estar compuestas por patrones de actividad (o procesamiento de la información) a varios niveles, y no solo en el caso de las que emplean los ordenadores. Veremos en el capítulo VI que la mente humana se puede considerar una máquina virtual (o más bien como un conjunto de máquinas virtuales que interactúan unas con otras, funcionando en paralelo y desarrolladas o aprendidas en momentos diferentes) que está instalada en el cerebro.

Para que la IA avance, es preciso que progrese la formulación de máquinas virtuales interesantes y útiles. Que haya más ordenadores físicamente potentes (más grandes, más rápidos) está muy bien. Puede que hasta sean necesarios para implementar en ellos ciertos tipos de máquinas virtuales, pero no se les puede sacar provecho hasta que las máquinas virtuales que se ejecuten sean informacionalmente potentes. (De la misma manera, para que haya avances en neurociencia se necesita comprender mejor qué máquinas virtuales psicológicas se ejecutan en las neuronas físicas: véase capítulo VII).

Se utilizan distintas clases de información del mundo exterior. Todo sistema de IA necesita dispositivos de entrada y de salida de datos, aunque sean solo un teclado y una pantalla. Suele haber también sensores con fines específicos (como cámaras o sensores táctiles con forma de bigotes electrónicos) y/o efectores (como sintetizadores de sonido para la música o el habla o manos robóticas). El programa de IA se conecta con estas interfaces del mundo informático, generando cambios en ellas, además de procesar información internamente.

Los procesos de IA suelen incluir también dispositivos internos de entrada y salida de datos que permiten a las distintas máquinas

virtuales que hay dentro del sistema interactuar entre sí. Por ejemplo, si una parte de un programa de ajedrez detecta una posible amenaza que sucede en otra, entonces puede conectarse con una tercera para buscar una táctica de bloqueo.

CATEGORÍAS PRINCIPALES DE LA IA

Cómo se procesa la información depende de la máquina virtual de que se trate. Como veremos en capítulos posteriores, hay cinco categorías principales y cada una de ellas tiene múltiples variaciones. Una es la IA clásica o simbólica, a veces llamada GOFAI (por las siglas en inglés de Good Old-Fashioned IA, IA a la antigua). Otra son las redes neuronales o el modelo conexionista. Además, están la programación evolutiva, los autómatas celulares y los sistemas dinámicos.

Cada investigador particular suele utilizar un solo método, aunque también existen las máquinas virtuales híbridas. Por ejemplo, en el capítulo IV se menciona una teoría sobre la actividad humana que fluctúa constantemente entre el procesamiento simbólico y el conexionista. (Esto explica cómo y por qué una persona que lleva a cabo una tarea planificada puede distraerse al notar en el medio algo que no guarda ninguna relación con su tarea). Y en el capítulo V se describe un dispositivo sensoriomotriz que combina robótica “situada”, redes neuronales y programación evolutiva. (Este dispositivo ayuda a que el robot encuentre el camino a “casa”, utilizando un triángulo de cartón como punto de referencia).

Además de sus aplicaciones prácticas, estos enfoques pueden ilustrar la mente, el comportamiento y la vida. Las redes neuronales son útiles para replicar elementos del cerebro y para el reconocimiento de patrones y aprendizajes. La IA clásica (especialmente cuando se combina con la estadística) puede replicar también el aprendizaje, la planificación y el razonamiento. La programación evolutiva esclarece la evolución biológica y el desarrollo cerebral. Los autómatas celulares y los sistemas

dinámicos se pueden utilizar para replicar el desarrollo de organismos vivos. Algunas metodologías se acercan más a la biología que a la psicología, y otras se aproximan más al comportamiento irreflexivo que al pensamiento deliberativo. Para entender por completo la gama de mentalidades nos harán falta todas y, seguramente, algunas más.

A muchos de los investigadores de IA no les interesa el funcionamiento de la mente: van detrás de la eficiencia tecnológica, no del entendimiento científico. Aunque sus técnicas se originaron en la psicología, ahora guardan escasa relación con ella. Veremos, no obstante, que para que progrese la IA con fines generales (la inteligencia artificial fuerte, IAF o AGI por sus siglas en inglés) hay que entender más profundamente la arquitectura computacional de la mente.

PREDICCIÓN DE LA IA

Lady Ada Lovelace predijo la IA en la década de 1840.[1] Más concretamente, predijo parte de ella. Al no haber atisbos de las redes neuronales ni de la IA evolutiva o dinámica, se centró en los símbolos y en la lógica. Tampoco sentía inclinación por el objeto psicológico de la IA, ya que su interés era puramente tecnológico.

Dijo, por ejemplo, que una máquina “podría componer piezas musicales elaboradas y científicas de cualquier grado de complejidad o extensión” y también que podría expresar “los grandes hechos de la naturaleza” y haría posible “una época gloriosa para la historia de las ciencias”. (Así que no le habría sorprendido ver que, dos siglos más tarde, los científicos utilizan “big data” y trucos de programación diseñados especialmente para mejorar los conocimientos de genética, farmacología, epidemiología… La lista es infinita).

La máquina que tenía en mente era la Máquina Analítica. Este dispositivo de engranajes y ruedas dentadas (que nunca se terminó de construir) lo había diseñado su gran amigo Charles Babbage en

1834. A pesar de estar concebida para el álgebra y los números, en lo esencial era equivalente a la computadora numérica polivalente.

Ada Lovelace reconoció la potencial generalidad de la Máquina, su capacidad para procesar símbolos que representasen “todas las materias del universo”. También describió varios fundamentos de la programación moderna: programas almacenados, subrutinas anidadas jerárquicamente, direccionamiento, microprogramas, estructuras de control, sentencias condicionales y hasta los bugs (errores de software). Pero no dijo nada sobre cómo podrían implementarse la composición musical o el razonamiento científico en la máquina de Babbage. La IA era posible, sí, pero cómo llegar a ella seguía siendo un misterio.

CÓMO EMPEZÓ LA IA

Aquel misterio lo aclaró un siglo después Alan Turing. En 1936, Turing demostró que, en principio, un sistema matemático que ahora se llama máquina universal de Turing[2] puede llevar a cabo todos los cálculos posibles. Este sistema imaginario crea y modifica combinaciones de símbolos binarios representados por “0” y “1”. Después de descifrar códigos en Bletchley Park durante la Segunda Guerra Mundial, Turing pasó el resto de la década de 1940 pensando en cómo podría hacer un modelo físico aproximado de su máquina definida de manera abstracta (contribuyó al diseño de la primera computadora moderna, que se terminó en Manchester en 1948) y en cómo se podía inducir a un artefacto semejante a desempeñarse con inteligencia.

A diferencia de Ada Lovelace, Turing aceptó ambos fines de la IA. Quería las nuevas máquinas para hacer cosas útiles que por lo general se supone que requieren inteligencia (quizá mediante técnicas muy antinaturales) y también para representar los procesos que acontecen en la mente de base biológica.

El objetivo primordial del artículo de 1950 en el que burlonamente planteó la prueba de Turing (véase capítulo VI) era

servir como manifiesto de la IA.[3] (Había escrito una versión más completa poco después de la guerra, pero la ley de Secretos Oficiales del Reino Unido impidió que se publicara). En él señalaba las cuestiones fundamentales del procesamiento de la información que intervienen en la inteligencia (juego, percepción, lenguaje y aprendizaje), y daba pistas sugerentes sobre lo que ya se había conseguido. (Solo "pistas", porque el trabajo que se hacía en Bletchley Park seguía siendo alto secreto). Incluso sugería planteamientos computacionales –como las redes neuronales y la computación evolutiva– que no fueron relevantes hasta mucho más tarde. Pero el misterio distaba mucho de aclararse. Eran observaciones muy generales: programáticas, no programas.

La convicción de Turing de que la IA debía ser posible de algún modo fue apoyada a principios de la década de 1940 por el neurólogo y psiquiatra Warren McCulloch y el matemático Walter Pitts en su artículo "A Logical Calculus of the Ideas Immanent in Nervous Activity" [Un cálculo lógico de las ideas inmanentes en la actividad nerviosa][4] en el que unieron el trabajo de Turing con otros dos elementos interesantes de principios del siglo xx: la lógica proposicional de Bertrand Russell y la teoría de las sinapsis neuronales de Charles Sherrington.

El aspecto fundamental de la lógica proposicional es que es binaria. Se supone que toda oración (llamada también proposición) es verdadera o falsa. No hay término medio; no se reconocen la incertidumbre o la probabilidad. Solo existen dos valores de verdad, esto es, verdadero y falso.

Además, para formar proposiciones complejas e inferir argumentos deductivos se utilizan conectivas lógicas (como y, o y si-entonces) cuyos significados se definen en función de la verdad / falsedad de las proposiciones que los componen. Por ejemplo, si dos (o más) proposiciones están conectadas por y, se supone que ambas / todas son verdaderas. Así, "Mary se casó con Tom y Flossie se casó con Peter" es verdadera si y solo si ambas "Mary se casó con Tom" y "Flossie se casó con Peter" son verdaderas. Si, de

hecho, Flossie no se casó con Peter, entonces la proposición compleja que contiene “y” es de por sí falsa.

McCulloch y Pitts podían juntar a Russell y Sherrington porque ambos habían descrito sistemas binarios. Se asignaron los valores verdadero / falso de la lógica a la actividad de encendido / apagado de las células cerebrales y a los 0 / 1 de cada estado de las máquinas de Turing. Sherrington no creía que las neuronas estuviesen estrictamente encendidas / apagadas, sino que también tenían umbrales fijos. Así, definieron las compuertas lógicas (los y, o y no computacionales) como redes neuronales minúsculas que podían interconectarse para representar proposiciones extremadamente complejas. Todo lo que pudiera expresarse mediante lógica proposicional podía ser calculado por alguna red neuronal y por alguna máquina de Turing.

En suma, la neurofisiología, la lógica y la computación se agruparon, y apareció también la psicología. McCulloch y Pitts creían (como muchos filósofos en aquel entonces) que el lenguaje natural se reduce, en lo esencial, a lógica. Por tanto, todo razonamiento y opinión, desde los argumentos científicos a los delirios esquizofrénicos, era agua para su molino teórico. Auguraron que, para el conjunto de la psicología, “una descripción detallada de la red [neural] aportaría todo lo que se puede conseguir en ese campo”.

La implicación principal estaba clara: uno y el mismo enfoque teórico (esto es, la computación de Turing) podía aplicarse a la inteligencia humana y a la artificial. (El artículo de McCulloch / Pitts influyó incluso en el diseño de los computadores. A John von Neumann, que pretendía en aquel entonces usar el sistema decimal, le hizo reflexionar y cambiar al código binario).

Turing, por supuesto, estaba de acuerdo. Pero no pudo aportar mucho más a la IA: la tecnología disponible era demasiado primitiva. A mediados de la década de 1950, sin embargo, se desarrollaron máquinas más potentes y/o más fáciles de usar. “Fácil de usar”, en

este caso, no significa que fuese más fácil pulsar los botones del ordenador o moverlo por la habitación. Más bien significa que era más fácil definir nuevas máquinas virtuales (por ejemplo, lenguajes de programación), que podrían utilizarse con mayor facilidad para definir máquinas virtuales de mayor nivel (por ejemplo, programas matemáticos o de planificación).

La investigación sobre la IA simbólica, en líneas generales con el mismo espíritu del manifiesto de Turing, comenzó a ambos lados del Atlántico. Un referente de finales de la década de 1950 fue el jugador de damas de Arthur Samuel, que llegó a los titulares de los periódicos porque aprendió a ganarle a su propio creador.[5] Era un indicio de que los ordenadores podrían desarrollar inteligencia suprahumana algún día y superar las capacidades de sus programadores.

El segundo de estos indicios tuvo lugar a finales de la década de 1950, cuando la Máquina de la Teoría Lógica no solo demostró dieciocho de los teoremas lógicos principales de Russell, sino que además halló una prueba más elegante para uno de ellos.[6] Fue algo verdaderamente impresionante, porque si bien Samuel no era más que un jugador de damas mediocre, Russell era un lógico de primera fila a nivel mundial. (Russell mismo estaba encantado con el logro, pero el Journal of Symbolic Logic rechazó publicar un artículo cuyo autor fuese un programa de ordenador, sobre todo porque no había demostrado un teorema nuevo).

La Máquina de la Teoría Lógica no tardó en ser superada por el Solucionador General de Problemas (GPS por sus siglas en inglés);[7] “superada” no porque pudiese sobrepasar a otros genios sobresalientes, sino porque no se limitaba a un solo campo. Como su nombre sugiere, el Solucionador General de Problemas podía aplicarse a cualquier problema que pudiera representarse (como se explica en el capítulo II) en términos de objetivos, subobjetivos, medios y operaciones. Les correspondía a los programadores identificar los objetivos, medios y operaciones relevantes para

cualquier campo específico, pero, una vez hecho esto, se le podía dejar el razonamiento al programa.

El GPS logró resolver el problema de “los misioneros y los caníbales”, por ejemplo. (Tres misioneros y tres caníbales en una ribera del río; una barca con espacio para dos personas; ¿cómo pueden cruzar el río todos sin que el número de caníbales sea mayor al de misioneros?). Es difícil incluso para los seres humanos, porque hay que retroceder para poder avanzar. (¡El lector puede intentar resolverlo usando moneditas!).

La Máquina de la Teoría Lógica y el Solucionador General de Problemas fueron ejemplos tempranos de la inteligencia artificial simbólica o GOFAI. Ahora están “pasados de moda”, por supuesto, pero también fueron “buenos”, como pioneros en el uso de heurísticas y planificación, ambas extremadamente importantes en la IA actual (véase capítulo II).

La inteligencia artificial simbólica no fue el único tipo de IA inspirada por el artículo “Logical Calculus”; también para el conexionismo fue alentadora. En la década de 1950, las redes de neuronas lógicas de McCulloch-Pitts, bien construidas expresamente, bien simuladas en computadores digitales, fueron usadas (por Albert Uttley, por ejemplo)[8] para crear modelos de aprendizaje asociativo y reflejos condicionados. (A diferencia de las redes neuronales actuales, hacían procesos locales, no distribuidos: véase el capítulo IV).

Pero los primeros modelos de redes no estaban dominados por completo por la neurología. Los sistemas que a mediados de la década de 1950 implementó Raymond Beurle (en ordenadores analógicos) eran muy diferentes.[9] En vez de redes de compuertas lógicas diseñadas al detalle, empezó con parrillas bidimensionales (matrices) de unidades conectadas al azar y umbrales variables. Vio que la autoorganización neuronal se debía a dinámicas de activación, construcción, expansión, persistencia, muerte y a veces de interacción.

Como bien notó Beurle, decir que una máquina basada únicamente en la lógica podía replicar los procesos psicológicos no era como decir que el cerebro fuese en realidad como tal máquina. McCulloch y Pitts ya lo habían señalado. Apenas cuatro años después de su revolucionario primer artículo, publicaron otro argumentando que el funcionamiento del cerebro se parece más a la termodinámica que a la lógica.[10] La lógica le cedió el paso a la estadística, la unidad a la colectividad y la pureza determinista al ruido de las probabilidades.

Dicho de otro modo, habían descrito lo que ahora se llama computación distribuida tolerante a fallos (véase capítulo IV). Entendieron este nuevo enfoque como "extensión" del anterior, no como contradicción, pero era más realista en el aspecto biológico.

CIBERNÉTICA

La influencia de McCulloch en los primeros tiempos de la IA no se limitó a la inteligencia artificial simbólica y al conexionismo. Sus conocimientos tanto de neurología como de lógica lo convirtieron en un líder inspirador para el incipiente movimiento cibernético de la década de 1940.

Los cibernéticos se centraron en la autoorganización biológica, que abarcaba varias clases de adaptación y metabolismo, incluyendo el pensamiento autónomo y el control motor, así como la regulación (neuro)fisiológica. Su concepto principal era la "causalidad circular" o retroalimentación, y la teleología u orientación a un fin era un concepto clave. Estas ideas estaban íntimamente relacionadas, ya que la retroalimentación dependía de las diferencias entre los objetivos: la distancia real al objetivo en cada momento servía de orientación para el siguiente paso.

Norbert Wiener (que diseñó misiles antibalísticos durante la guerra) dio nombre al movimiento en 1948; definiéndolo como "el estudio del control y la comunicación de animales y máquinas".[11]

Los cibernéticos que hacían modelos computacionales por lo general se inspiraban en la automatización y la computación análogica más que en la lógica y la computación digital. Sin embargo, la distinción no estaba bien definida. Por ejemplo, las diferencias entre objetivos se utilizaban tanto para controlar misiles teledirigidos como para enfocar la resolución de problemas simbólicos. Además, Turing, el campeón de la IA clásica, utilizaba ecuaciones dinámicas (en la descripción de la difusión química) para definir sistemas autoorganizados en los que la estructura nueva, como puntos o segmentos, podía surgir de cualquier procedencia homogénea (véase capítulo V).[12]

Otros de los primeros miembros del movimiento fueron el psicólogo experimental Kenneth Craik, el matemático John von Neumann, los neurólogos William Grey Walter y William Ross Ashby, el ingeniero Oliver Selfridge, el psiquiatra y antropólogo Gregory Bateson y el químico y psicólogo Gordon Pask.[13]

Craik, que murió en 1943 a los 31 años en un accidente de bicicleta, antes del advenimiento de las computadoras digitales, se refirió a la computación analógica al estudiar el sistema nervioso. Describió la percepción, la acción motriz y la inteligencia en general como si las orientase la retroalimentación de los “modelos” cerebrales.[14] Más tarde, su concepto de los modelos cerebrales o representaciones influiría enormemente en la IA.

A Von Neumann le había intrigado la autoorganización a lo largo de la década de 1930 y se quedó muy impresionado con el primer artículo de McCulloch y Pitts. Además de cambiar el diseño básico computacional, pasando del sistema decimal al binario, adaptó las ideas de ellos para explicar la evolución biológica y la reproducción. Dio forma a varios autómatas celulares: sistemas fabricados a partir de muchas unidades computacionales, cuyos cambios seguían reglas simples que dependían del estado de las unidades vecinas en cada momento.[15] Algunas podían replicar a otras. Llegó a definir un replicador universal capaz de copiar cualquier cosa, incluso a sí

mismo. Los errores en la replicación, decía, podían llevar a la evolución.

Von Neumann definió los autómatas celulares en términos informativos abstractos, aunque se podían encarnar de muchas formas, por ejemplo, en robots de auto-ensamblaje, en la difusión química de Turing, en las ondas físicas de Beurle o, como pronto se vio, en el ADN.

Desde finales de la década de 1940, Ashby desarrolló el Homeostato, un modelo electroquímico de homeostasis fisiológica.[16] Esta máquina misteriosa era capaz de adaptarse a un estado de equilibrio global sin importar los valores que se le asignaran en principio a sus 100 parámetros (aceptaba casi 400.000 condiciones iniciales diferentes). Demostraba la teoría de la adaptación dinámica de Ashby, tanto dentro del cuerpo (sin olvidar el cerebro) como entre el cuerpo y el medio exterior, en el aprendizaje por ensayo y error y en el comportamiento adaptativo.

Grey Walter también estaba estudiando el comportamiento adaptativo, pero de forma muy diferente.[17] Construyó mini-robots parecidos a tortugas, cuyo circuito electrónico sensoriomotor encajaba en la teoría de reflejos neuronales de Sherrington. Estos robots situados pioneros mostraban comportamientos naturales como seguir la luz, esquivar obstáculos y aprender de modo asociativo mediante reflejos condicionados. Eran lo bastante fascinantes como para ser mostrados al público en el Festival de Bretaña en 1951.

Diez años después, Selfridge (nieto del fundador de los grandes almacenes de Londres del mismo nombre) utilizó métodos simbólicos para implementar básicamente un sistema de procesamiento en paralelo llamado Pandemonium.[18]

Este programa de inteligencia artificial simbólica aprendía a reconocer patrones mediante muchos “demonios” de nivel inferior que buscaban cada uno una única entrada perceptual y transmitían

el resultado a demonios de un nivel superior. Sopesaban la consistencia de los rasgos reconocidos (por ejemplo, solo dos rayas horizontales en una F) y descartaban los rasgos que no cuadraban. Los niveles de fiabilidad podían variar y se tenían en cuenta: se hacía más caso a los demonios que gritaban más alto. Finalmente, un demonio principal escogía el patrón más plausible dada la evidencia disponible (muchas veces contradictoria). Este estudio influyó tanto en el conexionismo como en la IA simbólica. (Una ramificación muy reciente es el modelo de conciencia LIDA: véase el capítulo VI).

A Bateson le interesaban poco las máquinas, pero a finales de la década de 1960 basó sus teorías sobre cultura, alcoholismo y el “doble vínculo” en la esquizofrenia en unas ideas acerca de la comunicación (por ejemplo, la retroalimentación) que había sacado previamente de las reuniones de cibernéticos. Y desde mediados de la década de 1950, Pask (descrito por McCulloch como “el genio de los sistemas autoorganizados”) utilizó ideas cibernéticas y simbólicas en muchos proyectos diferentes, como teatro interactivo, robots que se comunicaban mediante música, arquitectura que aprendía y se adaptaba a los objetivos del usuario, conceptos que se autoorganizaban químicamente y máquinas de enseñar. Estas últimas permitían que se tomasen rutas diferentes a través de una representación compleja del conocimiento, así que servían tanto para el aprendizaje paso a paso como para los estilos cognitivos holísticos (y una variable tolerancia a lo irrelevante).

En suma, a finales de la década de 1960, y en algunos casos mucho antes, se estaban concibiendo e incluso implementando todos los tipos principales de IA.

La mayoría de los científicos que participaron son reverenciados de forma generalizada hoy. Pero Turing era el único invitado constante a los banquetes de la IA; durante muchos años, solo un subgrupo de la comunidad científica recordaba a los demás. Grey Walter y Ashby, en particular, cayeron prácticamente en el olvido hasta finales de 1980, cuando se los distinguió (igual que a Turing)

como abuelos de la vida artificial. Pask tuvo que esperar aún más para ser reconocido. Para entender el porqué, hay que saber cómo se dividieron los modelos computacionales.

CÓMO SE DIVIDIÓ LA IA

Hasta la década de 1960, no existía una distinción clara entre quienes hacían modelos del lenguaje o del razonamiento lógico y quienes hacían modelos del comportamiento motriz orientado a un fin o adaptativo. Algunos trabajaban en ambos. (Donald Mackay llegó a sugerir que se construyeran computadores híbridos capaces de combinar las redes neuronales con el procesamiento simbólico). Y todos simpatizaban entre sí. Los científicos que estudiaban la autorregulación fisiológica se veían tan embarcados en la misma empresa global como sus colegas orientados a la psicología. Todos asistían a las mismas reuniones: los seminarios interdisciplinares de Macy en Estados Unidos (dirigidos por McCulloch de 1946 a 1951) y la trascendental conferencia de Londres "La mecanización de los procesos mentales" (organizada por Uttley en 1958).[19]

A partir de la década de 1960, sin embargo, se desencadenó un cisma intelectual. A grandes rasgos, los interesados en la vida se quedaron en la cibernética y los interesados en la mente se volcaron en la computación simbólica. A los entusiastas de las redes les interesaba tanto el cerebro como la mente, por supuesto, pero estudiaban el aprendizaje asociativo en general, no los contextos semánticos específicos o el razonamiento, así que se metieron en la cibernética más que en la IA simbólica. Lamentablemente, había escaso respeto mutuo entre estos subgrupos cada vez más separados.

Era inevitable que surgieran camarillas sociológicas diferenciadas, ya que las cuestiones teóricas que se planteaban (biológicas –de varios tipos– y psicológicas –también de varios tipos–) eran diferentes, al igual que las competencias técnicas asociadas: a grandes rasgos, la lógica versus las ecuaciones diferenciales. La especialización creciente hizo que la comunicación

fuese cada vez más difícil y menos fructífera. Aquellas conferencias tan eclécticas quedaron en el olvido.

Aun así, la división no tendría que haber sido tan destemplada. La aprensión por parte de la cibernética y el conexionismo comenzó como una mezcla de envidia profesional e indignación justificada, motivadas por el enorme éxito inicial de la computación simbólica, por el interés periodístico que suscitó el evocador término “inteligencia artificial” (acuñado por John McCarthy en 1956 para nombrar lo que antes se había llamado “simulación computerizada”) y por la arrogancia (y el despliegue publicitario poco realista) que exhibieron algunos simbolistas.

Los miembros del campo simbólico eran menos hostiles en principio, porque se vieron como los ganadores de la competición de la IA. De hecho, ignoraron ampliamente la investigación sobre redes, a pesar de que algunos de sus líderes (Marvin Minsky, por ejemplo) habían empezado en ese campo.

En 1958, sin embargo, Frank Rosenblatt presentó una ambiciosa teoría de la neurodinámica en la que definía sistemas de procesamiento paralelo capaces de autorregular el aprendizaje a partir de una base aleatoria (y tolerantes al error por añadidura) y la implantó en parte en su máquina fotoeléctrica Perceptron.[20] A diferencia de Pandemonium, en Perceptron no era necesario que el programador analizase previamente los patrones de entrada. Los simbolistas no podían pasar por alto esta nueva forma del conexionismo, pero pronto la desestimaron con desprecio. Como se explica en el capítulo IV, Minsky (junto a Seymour Papert) lanzó una crítica hiriente en la década de 1960 alegando que los perceptrones son incapaces de computar algunas cosas básicas.[21]

A partir de entonces, los fondos para la investigación sobre redes neuronales se secaron. Este desenlace, que era el que buscaban deliberadamente los dos atacantes, profundizó los antagonismos dentro de la IA.

Para el público general, pareció entonces que la IA clásica era la única opción que había. Es cierto que las tortugas de Grey Walter habían recibido grandes elogios en el Festival de Bretaña. A finales de la década de 1950, hubo mucho despliegue en la prensa sobre el Perceptron de Rosenblatt y el reconocimiento de patrones del Adaline de Bernard Widrow (basado en el procesamiento de señales), pero la crítica de los simbolistas malogró aquel interés. La IA simbólica dominó los medios en las décadas de 1960 y 1970 (e influyó también en la filosofía de la mente).

Esta situación no duró. Las redes neuronales (como "sistemas PDP" haciendo procesamiento distribuido en paralelo) irrumpieron de nuevo en la esfera pública en 1986 (véase el capítulo IV). La mayoría de los profanos (y algunos iniciados, que tendrían que haberse dado cuenta antes) consideraron que este enfoque era del todo nuevo. Sedujo a los estudiantes de posgrado y despertó un enorme interés periodístico (y filosófico). Esta vez, los que fruncieron el ceño fueron los de la IA simbólica. Los PDP estaban de moda y se dijo que la IA clásica había fracasado casi por completo.

En cuanto a los demás cibernéticos, por fin salieron a la luz bajo la denominación vida artificial (A-Life) en 1987. Periodistas y estudiantes de posgrado los siguieron. Se volvió a cuestionar la IA simbólica.

En el siglo XXI, sin embargo, ha quedado claro que las preguntas diferentes requieren diferentes tipos de respuestas: cada mochuelo a su olivo. Aunque persisten rastros de la antigua animosidad, ahora hay sitio para el respeto e incluso para la cooperación entre los distintos enfoques. Por ejemplo, el "deep learning" o aprendizaje profundo se utiliza a veces en sistemas potentes que combinan la lógica simbólica con las redes probabilísticas multicapa; y otros enfoques híbridos incluyen ambiciosos modelos de consciencia (véase el capítulo VI).

Dada la rica variedad de máquinas virtuales que constituyen la mente humana, esto no debería sorprender tanto a nadie.

II

LA INTELIGENCIA GENERAL ES EL SANTO GRIAL

La IA de última generación es una cosa esplendorosa. Brinda en abundancia máquinas virtuales que realizan muchas clases diferentes de procesamiento de la información. No hay una llave maestra, no hay una técnica primordial que unifique el campo: los profesionales de la IA trabajan en temas muy diversos con muy poco en común en cuanto a objetivos y métodos. En este libro solo se pueden mencionar unos pocos de los avances recientes. En suma, el alcance metodológico de la IA es extraordinariamente amplio.

Se podría decir que ha tenido un éxito asombroso, ya que su alcance práctico, también, es extraordinariamente amplio. Existen multitud de aplicaciones de IA diseñadas para innumerables tareas específicas que utilizan en casi todos los campos de la vida legos y profesionales por igual. Muchas superan hasta a los seres humanos más expertos. En ese sentido, el progreso ha sido espectacular.

Pero los pioneros de la IA no solo aspiraban a sistemas especializados. También esperaban lograr sistemas con inteligencia general. Cada una de las capacidades humanas que replicaron (visión, razonamiento, lenguaje, aprendizaje y demás) ya conllevaba todo el cupo posible de desafíos. Además, se iban añadiendo cuando eran procedentes.

Según esos criterios, el progreso ha sido mucho menos impresionante. John McCarthy reconoció muy pronto lo necesario que era el “sentido común”[1] en la IA y habló de “generalidad en la inteligencia artificial” en los dos discursos de amplia repercusión que pronunció en los premios Turing, en 1971 y 1987, pero se estaba quejando, no celebrando. En 2016, sus quejas siguen sin respuesta.

El siglo XXI está viviendo un renacer del interés por la inteligencia artificial fuerte (IAF), motivado por el aumento reciente de la potencia de los ordenadores. Si se lograse la IAF, los sistemas de IA podrían depender menos de trucos de programación con un propósito específico y beneficiarse en su lugar de las facultades generales del razonamiento y la percepción, además del lenguaje, la creatividad y la emoción (de los que se habla en el capítulo III).

No obstante, es más fácil decirlo que hacerlo. La inteligencia general sigue siendo un reto fundamental pero muy esquivo. La IAF es el santo grial de este campo.

LOS SUPERORDENADORES NO BASTAN

Los superordenadores actuales son sin duda una ayuda para todo el que quiera hacer realidad este sueño. La explosión combinatoria –para la que se requieren más cálculos de los que se pueden realizar en realidad– ya no es la amenaza constante que era antes. Sin embargo, los problemas no se pueden resolver por el mero hecho de aumentar la potencia del ordenador.

En general, se necesitan nuevos métodos de resolución de problemas. Además, aunque un método en particular deba tener éxito en teoría, puede que necesite mucho tiempo y/o memoria para tener éxito en la práctica. Se dan tres ejemplos así (a propósito de las redes neuronales) en el capítulo IV. De la misma forma, una “solución” por fuerza bruta que enumerase todos los movimientos de ajedrez posibles requeriría más direcciones de memoria que electrones hay en el universo, así que ni siquiera bastaría con un montón de superordenadores.

La eficiencia también es importante: cuanto menor sea el número de cálculos, mejor. En resumen, hay que convertir los problemas en solubles y existen varias estrategias básicas para ello. La IA simbólica clásica les abrió el camino a todas y todas siguen siendo esenciales.

Una es concentrarse en una sola parte del espacio de búsqueda (la representación del problema del ordenador, en la que se supone que se encuentra la solución). Otra es crear un espacio de búsqueda menor

mediante supuestos simplificados. La tercera es ordenar la búsqueda con eficiencia. Otra más es crear un espacio de búsqueda distinto, representando el problema de una forma nueva.

Estas estrategias requieren heurísticas, planificación, simplificación matemática y representación del conocimiento, respectivamente. Las cinco secciones siguientes tratan de estas estrategias de la IA general.

BÚSQUEDA HEURÍSTICA

La palabra "heurística" tiene la misma raíz que "¡Eureka!": viene del griego encontrar o descubrir. Las heurísticas fueron las estrategias que eligió la inteligencia artificial simbólica inicial y se suelen considerar "trucos de programación". Pero el término no surgió con la programación; hacía mucho que lógicos y matemáticos lo conocían. En cuanto a la actividad humana de utilizar heurísticas para la resolución de problemas (ya sea conscientemente o no), se remonta a miles de años, mucho antes de que a Ada Lovelace se le pasara por la cabeza la IA.

Ya sea para seres humanos o máquinas, las heurísticas hacen que un problema sea más fácil de resolver. En IA, lo consiguen dirigiendo el programa hacia unas zonas determinadas del espacio de búsqueda y apartándolo de otras.

Muchas heurísticas, incluidas la mayoría de las que se utilizaron en los primeros tiempos de la IA, son reglas empíricas que no está garantizado que resulten. La solución puede estar en algún lugar del espacio de búsqueda que el sistema, por indicación de la heurística, ignora. Por ejemplo, "Protege a tu reina" es una regla muy útil en ajedrez, pero que a veces hay que desobedecer.

Se puede demostrar que otras heurísticas son adecuadas lógica o matemáticamente. Gran parte del trabajo actual sobre IA y ciencias de la computación tiene como objetivo identificar propiedades demostrables de los programas. Este es un aspecto de la "IA fácil de usar",[2] porque el uso de sistemas lógicamente poco fiables puede poner en peligro la seguridad humana (véase el capítulo VII). (No hay una distinción rigurosa entre heurísticas y algoritmos. Muchos algoritmos son, en efecto, miniprogramas que incluyen alguna heurística determinada).

Ya sean fiables o no, las heurísticas son un aspecto esencial de la investigación en la IA. La especialización creciente de la IA ya mencionada depende en parte de que se definan nuevas heurísticas que mejoren la eficiencia de manera espectacular, pero solo en un tipo de problema o espacio de búsqueda muy restringidos. Una heurística sumamente eficaz puede que no sea apta para que la "tomen prestada" otros programas de IA.

Dadas varias heurísticas, su orden de aplicación puede que sí sea importante. Por ejemplo, "Protege a tu reina" debería tenerse en cuenta antes de "Protege a tu alfil", aun cuando esta orden conduzca a veces al desastre. Órdenes distintas definirán árboles de búsqueda en el espacio de búsqueda. Definir y ordenar heurísticas son tareas cruciales para la IA moderna. (Las heurísticas también son muy importantes en la psicología cognitiva. El fascinante trabajo sobre "heurísticas rápidas y frugales", por ejemplo, demuestra cómo la evolución nos ha equipado con formas eficientes de reaccionar ante el medio).[3]

Las heurísticas buscan mediante fuerza bruta por todo el espacio de búsqueda superfluo, pero a veces se combinan con la búsqueda por fuerza bruta (limitada). El programa de ajedrez de IBM Deep Blue, que causó revuelo mundial al vencer al campeón Gary Kasparov en 1997, utilizaba chips de hardware dedicado que procesaban doscientos millones de posiciones por segundo para generar todos los movimientos posibles para las siguientes ocho jugadas.[4]

No obstante, tenía que usar heurísticas para seleccionar el "mejor" movimiento entre ellos. Y como sus heurísticas no eran fiables, ni siquiera Deep Blue fue capaz de vencer a Kasparov siempre.

PLANIFICACIÓN

La planificación, también, ocupa un lugar destacado en la IA actual, sobre todo en una amplia gama de actividades militares.[5] De hecho, el departamento de Defensa de Estados Unidos, que pagaba la mayoría de la investigación sobre IA hasta hace muy poco, ha dicho que el dinero ahorrado en logística (gracias a la planificación con IA) en el

campo de batalla en la primera guerra de Irak superó toda la inversión previa.

La planificación no se limita a la IA: todos la hacemos. Si el lector piensa en hacer la maleta para las vacaciones, por ejemplo, tiene que encontrar las cosas que se quiere llevar, que seguramente no estén todas en el mismo sitio. Quizá tenga que comprar algunas cosas (crema solar, tal vez). Debe decidir si juntar todas las cosas (quizá sobre la cama o una mesa) o si ponerlas en la maleta conforme las va encontrando. La decisión dependerá en parte de si quiere poner la ropa encima de todo para impedir que se arrugue. Le hará falta una mochila o una maleta o quizá dos: ¿cómo decidirlo?

Los programadores de inteligencia artificial simbólica que usaban la planificación como técnica tenían en cuenta este tipo de ejemplos pensados con todo detalle. (La IA basada en las redes neuronales es muy diferente, puesto que no intenta mimetizar la deliberación consciente: véase el capítulo IV). Esto es porque a los pioneros en IA responsables de la Máquina de la Teoría Lógica (véase el capítulo I) y el Solucionador General de Problemas les interesaba principalmente la psicología del razonamiento humano. Basaron sus programas en experimentos que habían hecho con sujetos humanos, a los que se les pidió que “pensaran en voz alta” para descubrir su propio proceso de pensamiento mientras hacían juegos de lógica.

Los planificadores de IA modernos no dependen tanto de las ideas obtenidas mediante la introspección consciente o la observación experimental. Y sus planes son mucho más complejos de lo que fueron en los primeros tiempos, pero la idea básica es la misma.

Un plan especifica una secuencia de acciones representadas a nivel general (un objetivo final, más subobjetivos y sub-subobjetivos…) para que no se tengan en cuenta todos los detalles a la vez. La planificación a un nivel adecuado de abstracción puede llevar a “podar” el espacio de búsqueda, para que algunos detalles no tengan que examinarse para nada. A veces, el objetivo final es un plan de acción en sí mismo, quizá programar envíos desde y a una fábrica o a un campo de batalla. Otras veces, es la respuesta a una pregunta; por ejemplo, un diagnóstico médico.

Para cualquier objetivo y situaciones previstas, el programa de planificación necesita: una lista de acciones (esto es, operadores simbólicos) o tipos de acciones (que se instan al completar los parámetros derivados del problema) que puedan realizar algún cambio relevante; para cada acción, un conjunto de requisitos previos necesarios (como agarrar algo que deberá estar al alcance) y heurísticas que den prioridad a los cambios necesarios y ordenen las acciones. Si el programa opta por una acción concreta, quizá tenga que crear un nuevo subobjetivo que cumpla con los requisitos previos. Este proceso de formulación de objetivos se puede repetir una y otra vez.

La planificación permite que el programa (y/o el usuario humano) averigüe si ya se han llevado a cabo las acciones y por qué. El "por qué" se refiere a la jerarquía de los objetivos: esta acción se llevó a cabo para cumplir con ese requisito previo, para alcanzar este y aquel subobjetivo. Los sistemas de IA, por lo general, usan técnicas de "encadenamiento hacia adelante" y "encadenamiento hacia atrás", que explican cómo encontró la solución el programa, lo que ayuda al usuario a juzgar si la acción / dictamen del programa es la apropiada.

Algunos planificadores actuales tienen decenas de miles de líneas de código que definen la jerarquía del espacio de búsqueda en numerosos niveles. Estos sistemas suelen ser significativamente distintos de los primeros planificadores.

Por ejemplo, en general no se supone que sea posible trabajar en todos los subobjetivos de manera independiente (es decir, que los problemas puedan descomponerse del todo). En la vida real, al fin y al cabo, el resultado de una acción orientada a un fin puede ser deshecho por otra. Los planificadores actuales son capaces de resolver problemas descomponibles parcialmente (trabajan por subobjetivos independientes), pero también de realizar procesos adicionales para combinar los subplanes derivados en caso necesario.

Los planificadores clásicos podían abordar solo problemas en los que el medio fuese del todo observable, determinista, finito y estático, pero algunos planificadores modernos son capaces de adaptarse a entornos que sean observables en parte (a saber, el modelo del sistema

del mundo puede ser incompleto y/o incorrecto) y probabilísticos. En esos casos, el sistema debe supervisar la evolución de la situación durante la ejecución con el fin de hacer cambios en el plan (y/o en sus propias “creencias” sobre el mundo) según proceda. Algunos planificadores modernos pueden hacer esto a lo largo de amplios periodos de tiempo: entablan continuamente la formulación, ejecución, adaptación y abandono de objetivos, según cambie el medio.

Se han añadido y se siguen añadiendo muchos otros avances a la planificación clásica.[6] Puede parecer sorprendente, pues, que algunos especialistas en robótica la rechazaran rotundamente en la década de 1980 y recomendasen en su lugar la robótica “situada” (véase el capítulo V). La noción de representación interna (de objetivos y acciones posibles, por ejemplo) también fue rechazada. Sin embargo, aquella crítica era errónea en su mayor parte. Los sistemas de los propios críticos puede que no representasen los objetivos, pero requerían representaciones de otras cosas, como de estimulaciones retinianas y recompensas. Además, incluso la robótica, origen de la crítica, por lo general necesita tanta planificación como respuestas puramente reactivas para construir robots que jueguen al fútbol, por ejemplo.[7]

SIMPLIFICACIÓN MATEMÁTICA

Mientras que las heurísticas dejan el espacio de búsqueda tal cual es (haciendo que el programa se fije solo en una parte), al simplificar las variables se crea un espacio de búsqueda irreal pero computacionalmente manejable.

Algunas de esas variables son matemáticas. Un ejemplo es la variable “i.i.d.” (independiente e idénticamente distribuida), que se suele utilizar en el aprendizaje automático y representa las probabilidades de los datos como si fueran mucho más simples de lo que son.

La ventaja de la simplificación matemática al definir el espacio de búsqueda es que se pueden usar métodos de búsqueda matemática; es decir, que el espacio se puede definir con claridad y es, al menos para los matemáticos, fácilmente inteligible. Eso no quiere decir que cualquier búsqueda matemática definida sea útil. Como ya se ha dicho, un método que matemáticamente garantice la resolución de todos los problemas

dentro de una clase determinada quizá no se pueda utilizar en la vida real porque necesitaría un tiempo infinito. Puede, sin embargo, sugerir estimaciones más factibles: véase el análisis sobre "retropropagación" en el capítulo IV.

Las simplificaciones no matemáticas de suposiciones en IA son legión y a menudo implícitas. Una es la suposición (tácita) de que los problemas se pueden definir y resolver sin tener en cuenta las emociones (véase el capítulo III). Muchas otras están incorporadas a la representación general del conocimiento que se utiliza cuando se especifica la tarea.

REPRESENTACIÓN DEL CONOCIMIENTO

Muchas veces, lo más difícil de la resolución de problemas en IA es presentarle el problema al sistema para que empiece a resolverlo. Aunque parezca que cualquiera puede comunicarse directamente con un programa (hablándole en inglés a Siri, quizá, o tecleando palabras en francés en el motor de búsqueda de Google), no es así. Ya se trate de textos o de imágenes, hay que presentarle al sistema la información ("conocimiento") de modo que pueda comprenderla; es decir, que pueda manejarla. (Si es entendimiento real o no se tratará en el capítulo VI).

Las formas que tiene la IA de hacer esto son muy diversas. Algunas son desarrollos o variaciones de los métodos generales de representación del conocimiento implantados por la inteligencia artificial simbólica. Otras, cada vez más, son métodos muy especializados, adaptados en cada caso a cuestiones específicas. Puede haber, por ejemplo, una nueva forma de representar imágenes de rayos X o fotografías de un tipo determinado de células cancerosas, diseñada especialmente para hacer posible un método muy específico de interpretación médica (es decir, que no sirve para hacerle un retrato al paciente).

Para conseguir la IAF, son fundamentales los métodos generales. Inspirados en los estudios psicológicos sobre cognición humana, incluyen series de SI-ENTONCES; representaciones de conceptos individuales; secuencias de acciones estereotipadas; redes semánticas e inferencias por lógica o probabilidad.

Veamos cada uno por separado. (Otra forma de representación del conocimiento, las redes neuronales, se describe en el capítulo IV).

PROGRAMAS BASADOS EN REGLAS

En la programación basada en reglas, un cúmulo de conocimiento / creencia se representa como una serie de sentencias condicionales que ligan condiciones con acciones: SI se cumple esta condición, ENTONCES realiza esta acción. Esta forma de representación del conocimiento se basa en la lógica formal (los sistemas de "producción" de Emil Post), pero los pioneros de la IA Allen Newell y Herbert Simon creían que subyacía en toda la psicología humana en general.

Tanto la condición como la acción pueden ser complejas y especificar una conjunción (o disyunción) de varios (o quizá muchos) aspectos. Si se cumplen varias condiciones al mismo tiempo, se da prioridad a la conjunción más inclusiva. Así, "SI el objetivo es cocinar paella" prevalecerá sobre "SI el objetivo es cocinar arroz" y si se añade a esa condición "con marisco", triunfará sobre la otra.

Los programas basados en reglas no especifican por adelantado el orden de los pasos. Más bien, cada regla espera que su Condición la desencadene. No obstante, este tipo de sistemas se puede utilizar para la planificación. Si no se pudiera, tendrían una utilidad muy limitada para la IA, aunque de diferente manera a una forma de programar anterior y más conocida (a veces llamada "función ejecutiva").

En programas con función ejecutiva (como el Solucionador General de Problemas y la Máquina de la Teoría Lógica: véase el capítulo I), la planificación se representa de manera explícita. El programador especifica una secuencia de instrucciones orientadas a un objetivo que hay que seguir paso a paso en un orden temporal estricto: "Hacer esto, luego hacer aquello; luego ir a ver si x es verdadero; si lo es, hacer esto y aquello; si no, hacer esto y aquello.

A veces, "esto" o "esto y aquello" es una instrucción explícita para establecer un objetivo o subobjetivo. Por ejemplo, a un robot con el

objetivo de salir de una habitación se le puede enseñar [sic] a que establezca el subobjetivo de abrir la puerta; luego, si al examinar el estado de la puerta comprueba que está cerrada, que establezca el sub-subobjetivo de agarrar el picaporte. (Un niño pequeño quizá necesite un sub-sub-sub-objetivo, como buscar a un adulto que agarre el picaporte inalcanzable y para eso quizá necesite varios objetivos de nivel aún menor).

Un programa basado en reglas también podría averiguar cómo escapar de la habitación. Sin embargo, la jerarquía del plan no se representaría como una secuencia de pasos explícitos ordenada temporalmente, sino como la estructura lógica implícita en el conjunto de sentencias condicionales SI-ENTONCES que comprende el sistema. Una Condición puede incluir el requerimiento de que se haya establecido antes cierto objetivo (SI quieres abrir la puerta y no eres lo bastante alto). Igualmente, una acción puede incluir que se establezca un nuevo objetivo o subobjetivo (ENTONCES pídeselo a un adulto). Los niveles menores se activarán de forma automática (SI quieres pedirle a alguien que haga algo, ENTONCES tienes que crear el objetivo de acercarte a él).

El programador, por supuesto, tiene que haber incluido las sentencias condicionales relevantes SI-ENTONCES (en nuestro ejemplo, las que traten de puertas y picaportes) pero no tiene que haber previsto todas las implicaciones lógicas potenciales de esas instrucciones. (Lo que al mismo tiempo es una ventaja y un inconveniente, porque durante mucho tiempo pueden pasar inadvertidas algunas inconsistencias potenciales).

Los objetivos / subobjetivos activos se registran en una “pizarra” central accesible para todo el sistema. La información que aparece en esa pizarra incluye no solo los objetivos activados sino también la información recibida y otros aspectos del proceso actual. (Esa idea ha influido en una destacada teoría neuropsicológica sobre la consciencia y en un modelo IA de consciencia basado en ella: véase el capítulo VI).

Los programas basados en reglas fueron de uso común en los primeros “sistemas expertos” de principios de la década de 1970, como MYCIN, que asesoraba a médicos humanos para identificar

enfermedades infecciosas y prescribir antibióticos, y DENDRAL, que realizaba análisis espectrales de moléculas de un ámbito particular de la química orgánica. MYCIN, por ejemplo, realizaba el diagnóstico médico equiparando síntomas y antecedentes de características corporales (condiciones) con conclusiones diagnósticas y/o sugería otras pruebas o medicación (acciones). Estos programas fueron el primer movimiento de la IA desde la esperanza de la generalidad a la práctica de la especialización; y el primer paso hacia el sueño de Ada Lovelace de la ciencia automática (véase el capítulo I).

La representación del conocimiento basada en reglas permite que los programas se vayan creando de manera gradual, conforme el programador (o quizá el mismo sistema de IAF) aprenda cosas sobre el dominio. Se pueden añadir reglas nuevas en cualquier momento, no hace falta reescribir el programa desde el principio. No obstante, hay una pega: si una regla nueva no guarda coherencia lógica con las ya existentes, el sistema no siempre hará lo que se supone que tiene que hacer; puede que ni siquiera se acerque a lo que se supone que tiene que hacer. Si se trata de un conjunto de instrucciones pequeño, estos conflictos lógicos se evitan con facilidad, pero en sistemas mayores son menos evidentes.

En la década de 1970, las nuevas sentencias lógicas provenían del diálogo constante con expertos humanos, a los que se les pedía que explicasen sus decisiones. Hoy, muchas de las reglas no provienen de la introspección consciente, pero son mucho más eficientes. Los sistemas expertos modernos (un término que ya casi ya no se usa) van desde programas enormes utilizados para investigaciones científicas y el comercio a humildes aplicaciones para móviles. Muchos superan a sus predecesores porque cuentan con formas adicionales de representación del conocimiento, como estadísticas, reconocimiento visual con fines específicos y/o el uso de big data (véase el capítulo IV).

Estos programas pueden asistir, o incluso reemplazar, a expertos humanos en campos muy restringidos. Algunos superan a los adalides mundiales en esos ámbitos. Hace casi cuarenta años, un sistema basado en reglas aventajó al experto supremo humano en diagnóstico de enfermedades de la soja.[8] Ahora hay ejemplos innumerables de programas que ayudan a los profesionales humanos de la ciencia, la

medicina, el derecho… y hasta del diseño de modas. (Lo que no es que sea una buena noticia precisamente: véase el capítulo VII).

MARCOS, VECTORES DE PALABRAS, SECUENCIAS, REDES SEMÁNTICAS

Otros métodos de representación del conocimiento utilizados normalmente atañen a conceptos individuales, no a dominios completos (como un diagnóstico médico o el diseño de un vestido).

Por ejemplo, se le puede decir a un ordenador lo que es una habitación especificando una estructura jerárquica de datos (también llamada “marco” o “frame”). Esto representa una habitación como algo que tiene suelo, techo, paredes, puertas, ventanas y muebles (cama, bañera, mesa de comedor, etcétera). Las habitaciones reales tienen un número variable de paredes, puertas y ventanas, así que en las “casillas” o “slots” del marco o frame se pueden incluir números específicos y también aportar atributos por defecto (cuatro paredes, una puerta, una ventana).

El ordenador puede usar tales estructuras de datos para encontrar analogías, responder preguntas, entablar una conversación o escribir o comprender una historia. Y son la base de CYC: un intento ambicioso (hay quien diría que demasiado ambicioso) de representar todo el conocimiento humano.

Sin embargo, los marcos pueden ser equívocos. Los atributos por defecto, por ejemplo, son problemáticos. (Algunas habitaciones no tienen ventanas y las habitaciones diáfanas no tienen puerta). Lo que es peor: ¿qué pasa con los conceptos cotidianos como tirar o derramar? La IA simbólica representa los conocimientos de sentido común de “física inexperta” mediante la construcción de marcos que codifican hechos como que un objeto físico se cae si no se sostiene, mientras que un globo de helio flota. Conseguir ser explícito en casos semejantes es una tarea interminable.

En algunas aplicaciones que usan técnicas recientes para manejar big data, un solo concepto se puede representar como un cúmulo o “nube” compuesto por cientos o miles de conceptos a veces

relacionados entre sí, y señalando las probabilidades de las muchas asociaciones que suelen aparecer juntas: véase el capítulo III. Igualmente, los conceptos se pueden representar mediante “vectores de palabras” en vez de con palabras. En este caso, el sistema de aprendizaje profundo descubre los rasgos semánticos que intervienen en conceptos diferentes, o conectan conceptos diferentes y los utiliza para predecir la siguiente palabra en la traducción automática, por ejemplo.[9] Sin embargo, estas representaciones no son todavía tan flexibles como los marcos clásicos para el razonamiento o la conversación.

Algunas estructuras de datos (llamadas “secuencias” o “scripts”) denotan secuencias de acciones conocidas.[10] Por ejemplo, meter a un niño en la cama normalmente conlleva arroparlo, contarle un cuento, cantarle una nana y apagar la luz. Este tipo de estructuras de datos se puede utilizar para preguntar, responder y también para sugerir preguntas. Si una madre omite apagar la luz, pueden surgir las cuestiones de ¿Por qué? y ¿Qué pasa después? Dicho de otro modo, en ello reside el origen de una historia. En consecuencia, esta forma de representación del conocimiento se utiliza en la escritura automática de relatos y sería necesaria para los ordenadores “de compañía” capaces de entablar una conversación humana normal (véase el capítulo III).

Una forma alternativa de representación del conocimiento de conceptos son las redes semánticas (que son redes localistas: véase el capítulo IV). Ideadas por Ross Quillian en la década de 1960 como modelos de memoria humana asociativa, entre ellas hay varios ejemplos de envergadura (por ejemplo, WordNet) que son ahora fuentes de datos públicas. Una red semántica une conceptos mediante relaciones semánticas como sinonimia, antonimia, subordinación, dominancia, parte-todo y muchas veces también incorpora a la semántica conocimientos fácticos mediante nexos asociativos (véase el capítulo III).

La red puede representar tanto palabras como conceptos, añadiendo enlaces que codifican sílabas, iniciales, fonética y homónimos. Una red así usa el JAPE de Kim Binsted y el STAND UP de Graeme Ritchie para crear chistes (de nueve tipos diferentes) basados en juegos de palabras, aliteraciones y cambio de sílabas. Por ejemplo: “P: ¿Qué clase de

criminal tiene fibra? R: Un asesino cereal”; “P: Hay dos peces en un tanque. Uno le pregunta al otro: ¿Sabes conducir esto?”.

Una advertencia: las redes semánticas no son lo mismo que las redes neuronales. Como se verá en el capítulo IV, las redes neuronales distribuidas representan el conocimiento de manera muy diferente. En ellas, los conceptos individuales no se representan con un solo nodo en una red asociativa definida minuciosamente, sino por la evolución de la actividad a través de una red completa. Los sistemas de este tipo toleran pruebas contradictorias, así que no les importunan los problemas de mantener la consistencia lógica (que se describirá en la siguiente sección), pero no pueden sacar conclusiones precisas. No obstante, son una clase de representación del conocimiento lo bastante importante (y una base para aplicaciones prácticas lo bastante importante) para merecer un capítulo aparte.

LA LÓGICA Y LA RED SEMÁNTICA

Si el objetivo final es la IAF, la lógica parece muy conveniente como representación del conocimiento, ya que es aplicable en general. En principio, se puede utilizar la misma representación (el mismo simbolismo lógico) para la visión, el aprendizaje, el lenguaje y demás, y para cualquier combinación entre ellos. Además, la lógica aporta métodos contundentes de demostración de teoremas para manejar la información.

Por eso, la clase de representación del conocimiento que se prefería en los inicios de la IA era el cálculo de predicados. Esta forma de la lógica tiene más capacidad de representación que la lógica proposicional, porque puede “meterse” en los enunciados para expresar su significado. Por ejemplo, examinemos el enunciado “Esta tienda tiene un sombrero para todo el mundo”. El cálculo de predicados puede distinguir con claridad estos tres significados posibles: “En esta tienda hay un sombrero diferente para cada persona”; “Existe en esta tienda un sombrero cuyo tamaño puede modificarse para que le quepa a cualquier persona”; y “En esta tienda existe un sombrero [suponemos que doblado] lo suficientemente grande como para meter en él a todo el mundo al mismo tiempo”.

Muchos investigadores de IA siguen prefiriendo la lógica de predicados. Los marcos de CYC, por ejemplo, se basan en la lógica de predicados, así como el procesamiento de lenguajes naturales (PLN o NLP, por sus siglas en inglés) y las representaciones de la semántica composicional (véase el capítulo III). A veces, la lógica de predicados se amplía con el objetivo de representar tiempo, causa, o deber / moral. Por supuesto, depende de si alguien ha desarrollado esas formas de la lógica modal, lo que no es fácil.

No obstante, la lógica también tiene sus desventajas. Una es la explosión combinatoria. El método de "resolución" para la demostración de teoremas lógicos, muy extendido en la IA, puede quedarse atascado sacando conclusiones que sean ciertas pero irrelevantes. Las heurísticas están para orientar y restringir las conclusiones y para decidir cuándo desistir (eso que el aprendiz de brujo no sabía hacer), pero no son infalibles.

Otra es que la demostración de teoremas mediante resolución supone que no no X implica X. Esta idea nos resulta conocida: en los argumentos por reducción del absurdo, se intenta encontrar una contradicción entre lo que alguien afirma y la premisa de la que parte. Si el campo sobre el que se está razonando es completamente comprensible, eso es lógicamente correcto; pero los usuarios de programas con resolución integrada (como muchos sistemas expertos) asumen por lo general que la imposibilidad de encontrar una contradicción implica que no existe ninguna contradicción (la así llamada "negación por fallo"). Suele ser un error. En la vida real, hay una gran diferencia entre demostrar que algo es falso y no poder demostrar que es verdadero (por ejemplo, si el lector se pregunta si su pareja le es o no infiel). Se debe a que se desconoce gran parte de la evidencia (premisas potenciales).

Una tercera desventaja es que en la lógica clásica ("monotónica"), una vez demostrado que algo es verdadero, sigue siendo verdadero. En la práctica no es siempre así. Se puede aceptar X como buena razón (quizá fuera una atribución predeterminada o incluso una conclusión a partir de un argumento detallado y/o evidencia), pero luego puede resultar que X ya no sea verdadero o no fuese verdadero desde el principio. En ese caso, se deben revisar las creencias propias de

manera acorde. En una representación del conocimiento basada en la lógica, es más fácil decirlo que hacerlo. Muchos investigadores, inspirándose en McCarthy,[11] han intentado desarrollar una lógica “no monotónica” capaz de tolerar valores de verdad cambiantes. De igual forma, se han formulado varias lógicas “difusas” en las que un enunciado se puede etiquetar como probable / improbable o desconocido, en vez de como verdadero / falso. A pesar de todo, no se ha encontrado ninguna defensa fiable contra la monotonicidad.

Los investigadores de IA que están desarrollando representaciones del conocimiento basadas en la lógica buscan cada vez más los átomos fundamentales de conocimiento o de significado en general. No son los primeros: McCarthy y Hayes lo hicieron en “Some Philosophical Problems from an AI Standpoint”[12] [Algunos problemas filosóficos desde el punto de vista de la inteligencia artificial]. Aquel artículo abordaba muchos dilemas conocidos, desde el libre albedrío a los contrafactuales, entre los que figuraban preguntas sobre la ontología básica del universo: estados, acontecimientos, propiedades, cambios, acciones... ¿qué?

A no ser que uno sea un metafísico de corazón (extraña pasión humana), ¿por qué deberían importarle estas preguntas? ¿Y por qué deberían plantearse “cada vez más” estas cuestiones? A grandes rasgos, la respuesta es que intentar diseñar IAF suscita preguntas sobre qué ontologías puede utilizar la representación del conocimiento. Estas preguntas también surgen al diseñar la web semántica.

La red semántica no es lo mismo que la World Wide Web en la que estamos desde la década de 1990, ya que la web semántica no es ni tecnología punta: es un estado del futuro. Si llega a existir y cuando exista, la búsqueda asociativa automática mejorará y se complementará con el aprendizaje automático, lo que permitirá que las aplicaciones y los navegadores accedan a la información desde cualquier parte de internet e incorporen de manera sensata diferentes elementos que razonen sobre asuntos. Es una tarea monumental. Además de necesitar enormes progresos técnicos en hardware y en infraestructura de las comunicaciones, este ambicioso proyecto (dirigido por Tim Berners-Lee) debe conseguir que los programas web-roaming (que navegan por internet) tengan una comprensión más profunda de lo que hacen.

Los motores de búsqueda como Google y los programas NLP en general son capaces de encontrar asociaciones entre palabras y/o textos, pero en ello no hay comprensión. En este caso, no se trata de una cuestión filosófica (para eso, véase el capítulo VI), sino empírica, y otro obstáculo para lograr la IAF. A pesar de algunos ejemplos tentadoramente engañosos (como WATSON, Siri y la traducción autómatica, de los que se habla en el capítulo III), los ordenadores actuales no captan el significado de lo que “leen” o “dicen”.

Un aspecto de esta falta de comprensión es la incapacidad de los programas para comunicarse unos con otros (aprender unos de otros), porque usan formas diferentes de representación del conocimiento y/o diferentes ontologías fundamentales. Si los desarrolladores de la web semántica pueden elaborar una ontología muy general, se podrá superar esta situación de torre de Babel. Por lo tanto, las cuestiones metafísicas sobre la IA suscitadas en la década de 1960 son importantes ahora por razones totalmente prácticas.

VISIÓN ARTIFICIAL

Los ordenadores actuales tampoco entienden las imágenes visuales como los seres humanos. (De nuevo, esta es una cuestión empírica: si las IAF pueden hacer fenomenología visual consciente se trata en el capítulo VI).

Desde 1980, las diversas representaciones del conocimiento utilizadas para la visión en IA se han basado mucho en la psicología, especialmente en las teorías de David Marr y James Gibson.[13] Marr se centró en construir representaciones 3D (invirtiendo el proceso de formación de la imagen), no en utilizarlas para la acción. Gibson sin embargo puso más énfasis en sus potencialidades visuales para la acción: pistas visuales que sugerían una ruta o una rama capaz de soportar peso o incluso un miembro de una especie amistosa u hostil. A pesar de la influencia de la psicología, los programas visuales de hoy tienen serias limitaciones.[14]

Es cierto que se han conseguido logros extraordinarios en visión artificial: el reconocimiento facial con un 98% de éxito, por ejemplo; o la lectura de letra cursiva manuscrita; o la observación de

comportamientos sospechosos en aparcamientos (como pararse continuamente delante de las puertas de los coches); o la identificación de ciertas células enfermas mejor que los patólogos humanos. Ante éxitos semejantes, no puede uno sino quedarse pasmado.

Pero los programas (muchos son redes neuronales: véase el capítulo IV) por lo general tienen que saber exactamente lo que están buscando: por ejemplo, una cara no al revés, no de perfil, no escondida parcialmente detrás de otra cosa y (para que haya un 98% de posibilidades de éxito) iluminada de cierta manera.

Lo de “por lo general” es importante. En 2012, el laboratorio de investigación de Google combinó mil grandes ordenadores (de 16 núcleos) para formar una red neuronal enorme, con más de mil millones de conexiones y equipada con aprendizaje profundo. Se le mostraron diez millones de imágenes aleatorias de vídeos de YouTube. No se le indicó qué buscar y las imágenes no estaban etiquetadas. Sin embargo, después de tres días, una UNI (una neurona artificial) había aprendido a reaccionar ante imágenes de caras de gatos y otra a rostros humanos.

¿Impresionante? Bueno, sí. Intrigante, también: los investigadores recordaron enseguida la idea de la “neurona abuela”. Desde la década de 1920, los neurocientíficos han discrepado sobre su existencia.[15] Decir que existe es decir que hay células en el cerebro (ya sean neuronas solas o pequeños grupos de neuronas) que se activan cuando y solo cuando perciben una abuela o algún otro elemento específico. Al parecer, lo mismo ocurría en la red de reconocimiento de gatos de Google. Y aunque las caras de los gatos tenían que estar completas y en la posición correcta, podían variar de tamaño o aparecer en posiciones diferentes dentro de un área de 200 x 200. Un estudio posterior, que capacitó al sistema con imágenes de rostros humanos seleccionadas de antemano cuidadosamente (y sin etiquetar), incluyendo algunos de perfil, dio como resultado una unidad que a veces (solo a veces) era capaz de distinguir rostros no orientados hacia el espectador.

Pronto habrá muchos más logros semejantes e incluso más impresionantes. Las redes multicapas han progresado mucho en reconocimiento facial y a veces son capaces de encontrar el elemento

más destacado de una imagen y generar una etiqueta para describirla (por ejemplo “gente comprando en un mercado al aire libre”).[16] El recientemente comenzado Large Scale Visual Recognition Challenge (reto de reconocimiento a gran escala) aumenta cada año el número de categorías visuales reconocibles y reduce las restricciones de las imágenes en cuestión (por ejemplo, el número y oclusión de objetos). No obstante, estos sistemas de aprendizaje profundo seguirán compartiendo algunas debilidades de sus predecesores.

Por ejemplo, no comprenderán (como el reconocedor facial de gatos) el espacio de tres dimensiones, no tendrán conocimiento de lo que es en realidad “perfil” u oclusión. Incluso los programas de visión artificial diseñados para robots solo ofrecen indicios sobre dichas cuestiones.

Los robots Mars Rover, como Opportunity y Curiosity (que aterrizaron en 2004 y 2012 respectivamente), dependen de trucos especiales para la representación del conocimiento: heurísticas hechas a medida para los problemas en 3D que se espera que afronten. Por lo general, no pueden explorar ni manipular objetos. Algunos robots simulan visión animada, en la que los propios movimientos del cuerpo les proporcionan información útil (porque cambian de entrada de información visual constantemente). Pero ni siquiera ellos pueden detectar una ruta posible o reconocer que pueden agarrar con la mano robótica ese objeto desconocido mientras que aquel otro no.

Cuando se publique este libro, puede que haya algunas excepciones, aunque también limitadas. Por ejemplo, los robots no comprenderán “No puedo coger eso”, porque no comprenderán puedo y no puedo. Esto es porque la lógica modal que se requiere probablemente aún no esté disponible para representar ese conocimiento.

A veces, la visión puede ignorar el espacio en tres dimensiones, al leer letra manuscrita, por ejemplo. Y en muchas tareas en dos dimensiones muy restringidas, la visión artificial comete menos errores que los seres humanos. De hecho, puede emplear técnicas nada naturales para analizar elementos detallados (en los rayos X, por ejemplo) que el ojo humano no podría reconocer. (De la misma forma, la visión artificial en 3D suele conseguir resultados notables con medios no naturales).

Pero hasta la visión artificial en 2D es limitada. A pesar del considerable esfuerzo de investigación sobre representaciones[17] analógicas o icónicas, la IA no puede utilizar de manera fiable diagramas para la solución de problemas, como hacemos nosotros para el razonamiento geométrico o cuando hacemos dibujitos abstractos en el reverso de un sobre. (De la misma forma, los psicólogos siguen sin entender cómo hacemos nosotros esas cosas).

En definitiva, la mayoría de los logros visuales humanos superan la IA actual. Los investigadores de IA por lo general no tienen claro qué preguntas plantear. Por ejemplo, pensemos en doblar con esmero un vestido de satén, que es un tejido escurridizo. Ningún robot puede hacerlo (aunque a algunos se les puede enseñar, paso a paso, cómo doblar una toalla rectangular). O consideremos cómo ponernos una camiseta: la cabeza entra primero y no por la manga, pero ¿por qué? Los problemas topológicos como estos casi nunca aparecen en la IA.

Nada de esto implica que la visión artificial a nivel humano sea imposible, pero conseguirlo es mucho más difícil de lo que cree la mayoría.

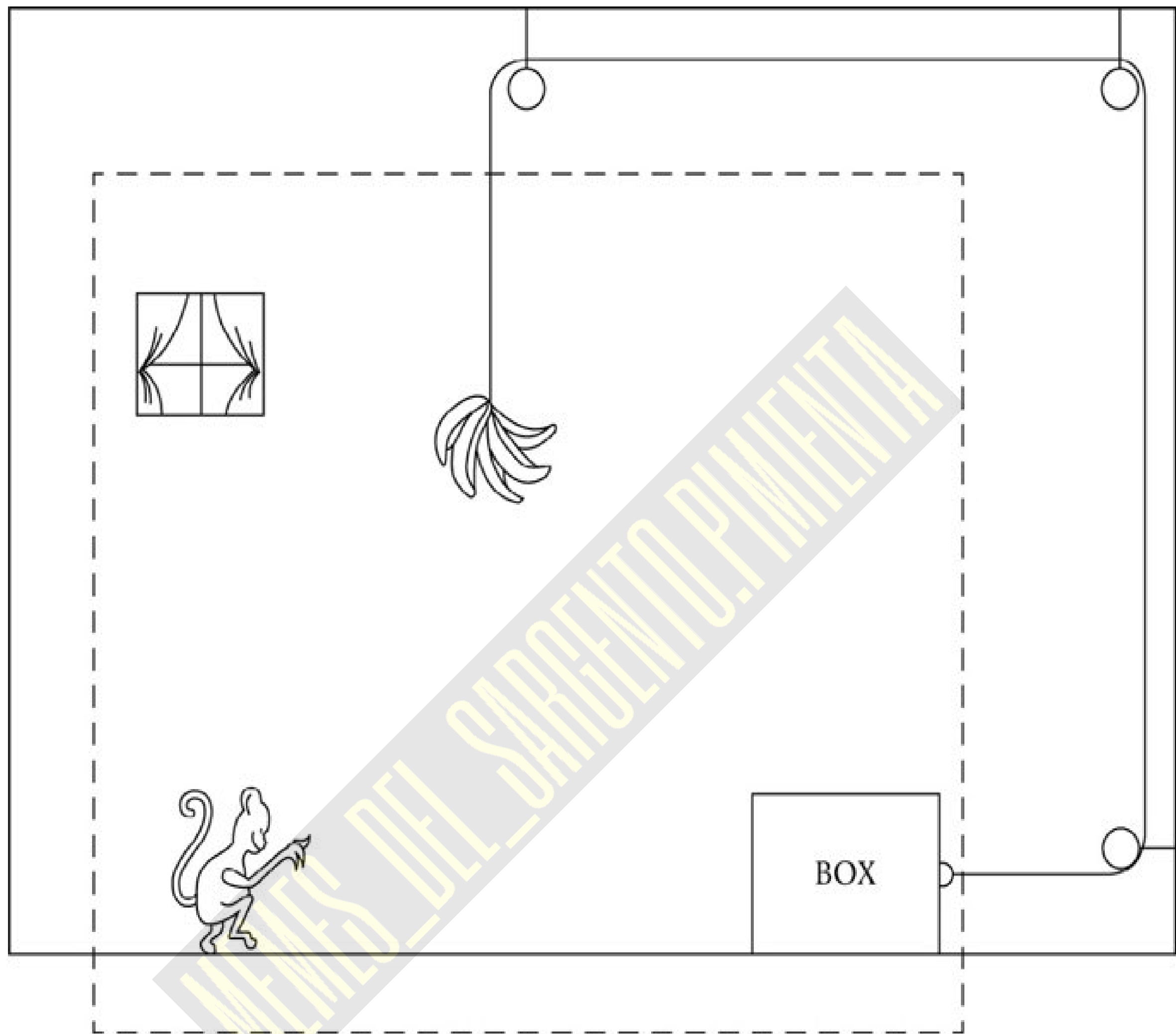
Esto se debe a que definirla es muy difícil. Así que este es un caso especial dentro del que se comentó en el capítulo I: que la IA nos ha enseñado que la mente humana es enormemente más rica y más sutil de lo que los psicólogos se imaginaban en un principio. De hecho, esta es la lección más importante que hay que aprender de la IA.

EL PROBLEMA DEL MARCO

Encontrar la representación del conocimiento apropiada, en cualquier campo, es difícil debido a la necesidad de evitar el problema del marco. (Cuidado: aunque este problema surge al usar marcos como representación del conocimiento de los conceptos, los significados de "marco" en este caso son diferentes).

Como definieron en un principio McCarthy y Hayes,[18] el problema del marco supone asumir (durante la planificación para robots) que una acción causará solo estos cambios, aunque podría causar también

aquellos. De manera más general, el problema del marco surge cuando el ordenador ignora las implicaciones (que los pensadores humanos asumen de manera tácita), porque no se han establecido explícitamente.



Cuadro 1. El problema del mono y los plátanos: ¿cómo consigue el mono los plátanos? (El enfoque usual de este problema asume, aunque no lo establece de forma explícita, que el mundo relevante es el mostrado dentro del marco de puntos. Dicho de otro modo, no existe nada fuera de este marco que cause cambios significativos en él si se mueve la caja.)

Reproducido de M. A. Boden, Artificial Intelligence and Natural Man (1977: 387).

El caso clásico es el problema del mono y los plátanos, en el que el solucionador de problemas (quizá un planificador de IA para un robot) asume que no existe nada relevante fuera del marco (véase la Figura 1).

Mi ejemplo favorito es: “Si un hombre de veinte años puede recoger cinco kilos de moras en una hora y una mujer de dieciocho puede recoger cuatro, ¿cuántos kilos de moras recogerán si van juntos?”. Está claro que “nueve” no es una respuesta plausible. Podrían ser muchos más (porque los dos intentarán lucirse) o es mucho más probable que sean muchos menos. Este ejemplo era todavía más revelador hace cincuenta años, la primera vez que supe de él. Pero ¿a qué se debe? ¿Qué tipos de conocimiento intervienen aquí? ¿Y podría resolver una IAF lo que parecen ser puros datos aritméticos?

El problema del marco surge porque los programas de IA no tienen el sentido humano de la relevancia (véase el capítulo III). Se puede evitar si se conocen todas las consecuencias posibles de todas las acciones posibles. En algunas ramas técnicas / científicas, es así. (Por eso los científicos de IA a veces afirman que el problema del marco está resuelto, o, si son muy prudentes, “más o menos” resuelto).[19] En general, sin embargo, no lo está. Esa es la razón principal por la que los sistemas de IA carecen de sentido común.

En suma, el problema del marco nos acecha por todos lados y es un obstáculo fundamental en la búsqueda de la IAF.

AGENTES Y COGNICIÓN DISTRIBUIDA

Un agente de IA es un procedimiento que se contiene a sí mismo (“autónomo”), comparable a veces a un movimiento reflejo y otras a una mente en miniatura.[20] Las aplicaciones de móviles o los correctores ortográficos se pueden llamar agentes, aunque por lo general no lo sean, porque los agentes normalmente cooperan. Usan su muy limitada inteligencia en cooperación con (o, en cualquier caso, junto a) otros para producir resultados que no pueden conseguir solos. La interacción entre agentes es tan importante como los agentes individuales.

Algunos sistemas agentes se organizan mediante control jerárquico: mandamases y mandados, por así decirlo, pero muchos ejemplifican la

cognición distribuida. Esto requiere de cooperación sin estructura jerárquica de mando (de ahí la evasiva, antes, entre “en cooperación con” y “junto a”). No hay plan central, no hay influencia impuesta desde arriba y no hay ningún individuo que posea todo el conocimiento.

Las hileras de hormigas, la navegación marítima y la mente humana son ejemplos naturales de cognición distribuida. Las hileras de hormigas las genera el comportamiento de muchas hormigas individuales que al andar dejan caer (y siguen) sustancias químicas de manera automática. De la misma forma, la navegación y las maniobras de los barcos resultan del engranaje de las actividades de mucha gente: ni siquiera el capitán tiene todos los conocimientos necesarios y algunos miembros de la tripulación tienen muy pocos. Hasta una sola mente precisa de cognición distribuida, ya que incluye muchos subsistemas cognitivos, motivacionales y emocionales (véanse los capítulos IV y VI).

Ejemplos artificiales son las redes neuronales (véase el capítulo IV); el simulador informático de navegación marítima de un antropólogo[21] y el trabajo sobre vida artificial de la robótica situada, la inteligencia de enjambre y la robótica de enjambres (véase el capítulo V); los modelos de IA simbólica de los mercados financieros (siendo los agentes los bancos, los fondos de inversión y los grandes accionistas) y el modelo de consciencia LIDA (véase el capítulo VI).

El conocimiento sobre cognición distribuida también ayuda a diseñar la interacción persona-ordenador como espacio de trabajo colaborativo e interfaces de ordenadores; esto se debe a que (como dice Yvonne Rogers) deja claras “las complejas interdependencias entre las personas, artefactos y sistemas tecnológicos que por lo general se pueden pasar por alto al utilizar teorías cognitivas tradicionales”.

Queda claro, entonces, que la IAF a nivel humano requeriría de cognición distribuida.

APRENDIZAJE AUTOMÁTICO

La IAF a nivel humano incluiría también el aprendizaje automático,[22] aunque no tendría por qué ser igual al humano. Este campo se originó en el trabajo de los psicólogos sobre aprendizaje y refuerzo de

conceptos. Sin embargo, ahora depende de técnicas aterradoramente matemáticas, porque las representaciones del conocimiento utilizadas requieren teoría de probabilidades y estadística. (Se podría decir que la psicología ha quedado muy atrás. Desde luego, algunos sistemas modernos de aprendizaje automático guardan poca o ninguna semejanza con lo que es verosímil que pase en la mente humana. Sin embargo, el uso creciente de la probabilidad bayesiana en este campo de la IA establece un paralelo con teorías recientes de la psicología cognitiva y la neurociencia).

El aprendizaje automático actual es muy lucrativo. Se utiliza para la minería de datos y, como los superordenadores realizan un millón de billones de cálculos por segundo, para procesar big data (véase el capítulo III).

Algunos aprendizajes automáticos usan redes neuronales, aunque muchos recurren a la IA simbólica complementada con potentes algoritmos estadísticos. De hecho, la estadística es la que de verdad hace el trabajo; la inteligencia artificial simbólica se limita a guiar al trabajador a su lugar de trabajo. En consecuencia, algunos profesionales consideran el aprendizaje automático como ciencias de la computación y/o estadística, no como IA. Sin embargo, no hay límites claros. (Algunos científicos computacionales rechazaron a propósito la etiqueta "IA" de McCarthy por sus problemáticas implicaciones filosóficas: véase el capítulo VI. Y otros la evitan porque no aprueban la naturaleza experimental, relativamente poco sistemática, de la mayoría [aunque desde luego no todas] de las investigaciones sobre IA).

El aprendizaje automático tiene tres grandes categorías: supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. (Estas distinciones se originaron en la psicología y quizá impliquen diferentes mecanismos neurofisiológicos; el aprendizaje por refuerzo, en todas las especies animales, requiere de dopamina).

En el aprendizaje supervisado, el programador "entrena" al sistema definiendo un conjunto de resultados deseados para una serie de entradas (ejemplos con etiqueta y ejemplos no relevantes) y notificando si se han obtenido. El sistema de aprendizaje genera hipótesis sobre los elementos relevantes. Cuando clasifica incorrectamente, corrige las

hipótesis en consonancia. Los mensajes de error específicos son cruciales (no la mera notificación de que se ha cometido uno).

En el aprendizaje no supervisado, el usuario proporciona los resultados no deseados o mensajes de error. El aprendizaje se rige por el principio de que, cuando hay elementos que concurren, generan la expectativa de que vuelvan a concurrir en el futuro. El aprendizaje no supervisado puede utilizarse para descubrir conocimiento. Los programadores no necesitan saber qué patrones / agrupaciones existen en los datos: el sistema los encuentra por sí mismo.

Por último, el aprendizaje por refuerzo se rige por principios análogos al castigo y la recompensa: mensajes de notificación que le dicen al sistema si lo que acaba de hacer está bien o mal. En refuerzo no es solo binario por lo general: también se representa con números, como la puntuación de un videojuego.

“Lo que acaba de hacer” puede ser una sola decisión (como un movimiento en un juego), o una serie de decisiones (por ejemplo, unos movimientos de ajedrez que terminan en jaque mate). En algunos videojuegos, la puntuación numérica se actualiza con cada movimiento. En situaciones muy complejas, como en el ajedrez, el éxito (o el fracaso) se señala solo después de muchas decisiones y algún procedimiento de asignación de crédito identifica las decisiones que es más probable que conduzcan al éxito. (La IA evolutiva es una forma de aprendizaje por refuerzo en la que el éxito lo controla la función de aptitud: véase el capítulo V).

El aprendizaje automático simbólico asume en general (lo que obviamente no es cierto) que la representación del conocimiento para el aprendizaje tiene que incluir alguna forma de distribución de probabilidad. Y muchos algoritmos de aprendizaje asumen (lo que suele ser falso) que todas las variables de los datos tienen la misma distribución de probabilidad y que son todos mutuamente independientes. Esto es porque este supuesto i.i.d. (independiente e idénticamente distribuido) se sustenta en muchas teorías matemáticas sobre probabilidad, en las que se basan los algoritmos. Los matemáticos adoptaron las variables i.i.d. porque simplificaban las matemáticas. De la

misma forma, utilizar las i.i.d. en IA simplifica el ámbito de búsqueda, y por lo tanto facilita la solución de problemas.

La estadística bayesiana, sin embargo, se ocupa de las probabilidades condicionales, en las que los objetos / hechos no son independientes. En este caso, la probabilidad depende de la distribución de las muestras por el dominio. Además de ser más realista, esta forma de representación del conocimiento permite que cambien las probabilidades si aparecen muestras nuevas. Las técnicas bayesianas son cada vez más importantes en IA y en psicología y neurociencia también. Las teorías sobre “el cerebro bayesiano” (véase el capítulo IV) sacan partido del uso de muestras no i.i.d. para dirigir y afinar el aprendizaje no supervisado en la percepción y el control motor.

Como hay varias teorías sobre probabilidad, hay muchos algoritmos apropiados diferentes para distintas clases de aprendizaje y conjuntos diferentes de datos. Por ejemplo, las máquinas de vectores de soporte (que aceptan variables i.i.d.) son de uso común en el aprendizaje supervisado, en especial si el usuario carece de conocimiento especializado previo sobre el entorno. Los algoritmos “bolsa de palabras” son útiles cuando se puede ignorar el orden de los elementos (como en la búsqueda de palabras, no de frases). Y, si se suprimen las variables i.i.d., las técnicas bayesianas (“máquina Helmholtz”) son capaces aprender a partir de la distribución muestral.

La mayoría de los profesionales del aprendizaje automático usan métodos estadísticos disponibles en el mercado. Los creadores de tales métodos son muy preciados por la industria: Facebook ha contratado recientemente al creador de las máquinas de vectores de soporte y, en 2013-2014, Google contrató a varios de los principales impulsores del deep learning o aprendizaje profundo.[23]

El aprendizaje profundo es un nuevo avance muy prometedor basado en redes multicapa (véase el capítulo IV) mediante el que se reconocen patrones en los datos de entrada en diferentes niveles jerárquicos.[24] Dicho de otro modo, el aprendizaje profundo descubre una representación del conocimiento con varios niveles, por ejemplo, píxeles para detectores de contraste, para detectores de bordes, para detectores de formas, para partes de objetos, o para objetos.

Un ejemplo es el detector facial de gatos que surgió de la búsqueda de Google en YouTube. Otro, publicado recientemente en la revista Nature, es un aprendiz por refuerzo (el algoritmo “DQN”) que ha aprendido a jugar a los clásicos de la Atari 2600 2D.[25] Pese a que se le dan como entradas solo píxeles y puntuaciones de los juegos (y conociendo previamente solo el número de acciones disponibles para cada juego), supera al 75% de los seres humanos en 29 de los 49 juegos y obtiene mejores resultados que los testeadores profesionales en 22.

Queda por ver hasta dónde puede llegar este logro. Aunque a veces DQN encuentre la estrategia óptima mediante acciones ordenadas en el tiempo, no es capaz de dominar los juegos cuya planificación abarque un periodo más largo de tiempo.

La futura neurociencia podría proponer mejoras a este sistema. La versión actual se inspira en el receptor de visión de Hubel y Wiesel (células del córtex visual que reaccionan solo al movimiento o solo a líneas con una orientación particular. No es nada del otro mundo, los receptores de Hubel y Wiesel también inspiraron Pandemonium: véase el capítulo I). Excepcionalmente, esta versión de DQN se inspira también en la “reproducción de la experiencia” que tiene lugar en el hipocampo durante el sueño. Igual que el hipocampo, el sistema DQN almacena un conjunto de muestras o experiencias pasadas y las reactiva rápidamente durante el aprendizaje. Esta característica es crucial: los diseñadores señalaron “un deterioro grave” del rendimiento cuando se desactivaba.

SISTEMAS GENERALISTAS

El jugador de Atari causó revuelo (y mereció ser publicado en Nature) en parte porque parecía un paso hacia la IAF. Un solo algoritmo, sin usar representaciones del conocimiento manufacturadas, aprendió un amplio rango de competencias en una serie de tareas que requerían entradas sensoriales de dimensión relativamente alta. Ningún programa anterior había hecho eso.

Sin embargo, (como se señaló al principio de este capítulo), la IAF completa haría mucho más. Por difícil que resulte construir una IA especialista de alto rendimiento, construir una IA generalista es varios órdenes de magnitud más difícil. (El aprendizaje profundo no es la solución: sus aficionados admiten que “se necesitan paradigmas nuevos” para combinarlo con el razonamiento complejo, que en el código académico significa “no tenemos ni idea”).[26] Esa es la razón por la que la mayoría de los investigadores de IA perdieron aquella esperanza inicial, recurriendo en su lugar a tareas variopintas muy definidas, a menudo con un éxito espectacular.

Entre los pioneros de la IAF que conservaron su ambiciosa esperanza estaban Newell y John Anderson. Idearon SOAR y ACT-R respectivamente: sistemas comenzados a principios de la década de 1980 que siguen en desarrollo (y utilizándose) unas tres décadas después. Sin embargo, simplificaron demasiado la tarea, enfocándose en un solo subconjunto de competencias humanas.

En 1962, el colega de Newell, Simon, había observado el camino zigzagueante de una hormiga sobre un terreno accidentado. Cada movimiento, dijo, es una reacción directa a la situación que percibe la hormiga en ese momento (esta es la idea principal de la robótica situada: véase el capítulo V). Diez años después, el libro de Newell y Simon Human Problem Solving describía nuestra inteligencia como algo similar.[27] Según su teoría psicológica, la percepción y la acción motora se complementan con representaciones internas (sentencias condicionales o “producciones”) guardadas en la memoria o recién creadas durante la resolución de problemas.

“Los seres humanos, vistos como sistemas de comportamiento –dijeron–, son bastante sencillos”. Pero las complejidades conductuales son significativas. Por ejemplo, demostraron que un sistema con solo catorce sentencias condicionales SI-ENTONCES podía resolver problemas criptoaritméticos (por ejemplo, aplica las letras a los dígitos 0 a 9 en esta suma: DONALD + GERALD = ROBERT, donde D = 5). Algunas condicionales se ocupan de la organización de objetivos / subobjetivos, otras dirigen la atención (a una letra o columna específica), otras recuperan pasos previos (resultados intermedios),

otras reconocen intentos fallidos y otras retroceden para recuperarse de ellos.

La criptoaritmética, alegaron, ejemplifica la arquitectura computacional de todo comportamiento inteligente, así que este enfoque psicológico encajaba en una IA generalista. A partir de 1980, Newell (junto a John Laird y Paul Rosenbloom) desarrolló SOAR (Success Oriented Achievement Realized), concibiéndolo como un modelo de la cognición como un todo.[28] Su razonamiento aunaba percepción, atención, memoria, asociación, inferencia, analogía y aprendizaje. Las respuestas (situadas) como las de las hormigas se combinaron con deliberación interna. De hecho, la deliberación solía dar como resultado respuestas reflejas, porque una secuencia de subobjetivos utilizada previamente podía “condensarse” en una sola norma.

De hecho, SOAR no consiguió replicar todos los aspectos de la cognición y más tarde se amplió cuando se reconocieron algunas de las carencias. La versión actual se utiliza para muchos propósitos, desde el diagnóstico médico a la programación de tareas industriales.

La familia ACT-R (Adaptive Control of Thought o control adaptativo de pensamiento) de Anderson son sistemas híbridos (véase el capítulo IV) desarrollados mediante la combinación de sistemas de producción y redes semánticas.[29] Estos programas, que reconocen las probabilidades estadísticas en el entorno, replican la memoria asociativa, el reconocimiento de patrones, el significado, el lenguaje, la solución de problemas, el aprendizaje, la imaginería y (desde 2005) el control perceptuo-motor. ACT-R es, sobre todo, un ejercicio de IA científica. Mientras que el aprendizaje automático comercial ha olvidado sus raíces psicológicas, ACT-R sigue profundizando en ellas (recientemente ha incluido también la neurociencia: por ejemplo, conjuntos de sentencias condicionales SI-ENTONCES en paralelo a sistemas cerebrales “modulares”).

Un elemento clave de ACT-R es la integración de conocimiento del procedimiento y la enunciación. Alguien puede saber que un teorema de Euclides es verdadero sin saber cómo usarlo en una demostración geométrica. ACT-R puede aprender a aplicar una verdad proposicional mediante la creación de cientos de nuevas producciones que controlen

su uso en muchas circunstancias diferentes. Aprende qué objetivos, subobjetivos y sub-subobjetivos son relevantes en qué condiciones y qué resultados tendrá una acción en particular en diferentes circunstancias. Y (como SOAR) puede condensar en una sola regla varias reglas que normalmente se lleven a cabo secuencialmente. Es un paralelismo de los modos diferentes en que resuelven "el mismo" problema un experto humano y un novato: sin pararse a pensarlo o con laboriosidad.

ACT-R tiene aplicaciones diversas. Sus tutores matemáticos ofrecen retroalimentación personalizada, lo que pasa por un conocimiento relevante del dominio y de la estructura de objetivos / subobjetivos de resolución de problemas. Gracias a que condensa varias reglas en una sola, el tamaño del grano de sus sugerencias cambia a medida que el aprendizaje del estudiante va avanzando. Y hay otras aplicaciones relacionadas con el procesamiento de lenguajes naturales: la interacción persona-ordenador; la memoria humana y la atención; conducir y volar o la búsqueda visual en internet.

SOAR y ACT fueron contemporáneos de otro intento temprano de IAF: el CYC de Douglas Lenat. Este sistema de IA simbólica se lanzó en 1984 y sigue en desarrollo continuo.[30]

En 2015, CYC contenía 62.000 "relaciones" capaces de asociar conceptos de su base de datos y millones de nexos entre esos conceptos, incluyendo asociaciones semánticas y factuales almacenadas en grandes redes semánticas (véase el capítulo III) y hechos innumerables de física básica (el conocimiento sin formalizar de los fenómenos físicos –como dejar caer y derramar– que tienen todos los seres humanos). El sistema usa lógica monotónica y no monotónica y también probabilidades para razonar sobre sus datos. (En este momento, todos los conceptos y relaciones se codifican manualmente, pero se está añadiendo aprendizaje bayesiano, lo que permitirá a CYC aprender de internet).

Lo han usado varias agencias gubernamentales de Estados Unidos, departamento de Defensa incluido (para vigilar grupos terroristas, por ejemplo), los institutos nacionales de salud y algunos bancos importantes y compañías de seguros. Se ha hecho pública una versión

reducida (OpenCyc) como fuente de información para diversas aplicaciones y está disponible un compendio más completo (ResearchCyc) para los que trabajan en IA. Aunque OpenCyc se actualiza con regularidad (la última vez fue en 2014), contiene solo un pequeño subconjunto de la base de datos de CYC y un pequeño subconjunto de reglas de inferencia. Con el tiempo, el sistema completo (o casi completo) estará disponible en el mercado. Sin embargo, podría caer en malas manos, a menos que se tomen medidas específicas para evitarlo (véase el capítulo VII).

Lenat describió CYC en IA Magazine (1986) como “Usar el conocimiento de sentido común para superar la fragilidad y los obstáculos en la adquisición de conocimientos”. Es decir, se estaba refiriendo específicamente al desafío profético de McCarthy. Hoy es el líder de los modelos del razonamiento de “sentido común” y también de la “comprensión” de los conceptos con los que trata (algo que al parecer no pueden hacer ni los programas NLP más impresionantes: véase el capítulo III).

No obstante, tiene muchas debilidades. Por ejemplo, no se las arregla bien con la metáfora (aunque la base de datos incluye muchas metáforas muertas, claro). Ignora varios aspectos de la física básica. Su NLP, aunque siempre está mejorándose, es muy limitado, y todavía no incluye la visión. En suma, a pesar de sus objetivos enciclopédicos, no engloba todo el conocimiento humano.

EL SUEÑO REVITALIZADO

Newell, Anderson y Lenat han estado en segundo plano, trabajando como hormiguitas, durante treinta años. Últimamente, sin embargo, el interés por la IAF se ha avivado notablemente. En 2008 se estableció una conferencia anual y a SOAR, ACT-R y CYC se le están uniendo otros sistemas supuestamente generalistas.

Por ejemplo, en 2010, Tom Mitchell, pionero en aprendizaje automático, lanzó NELL (Never-Ending Language Learner) de la universidad Carnegie Mellon. Este sistema de “sentido común” forma su conocimiento rastreando la web sin parar (lleva cinco años, al momento de escribir este libro) y aceptando correcciones en línea del público. Es

capaz de hacer inferencias sencillas basadas en sus datos (sin etiquetar); por ejemplo, el deportista Fulanito juega al tenis, ya que está en el equipo de la copa Davis. Partiendo de una ontología de doscientas categorías y relaciones (por ejemplo, maestro, es debido a), después de cinco años había ampliado la ontología y acumulado noventa millones de creencias candidatas, cada una con su propio nivel de confianza.

Las malas noticias son que NELL no sabe, por ejemplo, que se puede usar una cuerda para tirar de objetos, pero no para empujarlos. De hecho, el sentido común putativo de todos los sistemas de IAF es sumamente limitado. Las afirmaciones de que el famoso problema del marco ha quedado "resuelto" son muy engañosas.

NELL ahora tiene un programa hermano, NEIL (Never-Ending Image Learner), en el que unas IAF en parte visuales combinan una representación del conocimiento lógico-simbólica con representaciones analógicas o gráficas (una distinción hecha hace años por Aaron Sloman, pero que todavía no se entiende muy bien).

Asimismo, CALO (Cognitive Assistant that Learns and Organizes) del Stanford Research Institute produjo un producto derivado, la aplicación Siri (véase el capítulo III), que Apple compró en 2009 por doscientos millones de dólares. En la actualidad hay proyectos activos comparables, como el fascinante LIDA de Stan Franklin (del que se habla en el capítulo VI) y OpenCog de Ben Goertzel, que aprende datos y conceptos dentro de un rico mundo virtual y también de otros sistemas de IAF (LIDA es uno de los dos sistemas generalistas enfocados en la consciencia, el otro es CLARION).[31]

Un proyecto aún más reciente de IAF, iniciado en 2014, pretende desarrollar "una arquitectura computacional para la capacidad moral de los robots" (véase el capítulo VII). Además de las dificultades mencionadas antes, tendrá que enfrentarse a los muchos problemas referentes a la moral. Un sistema genuinamente humano no haría menos. No es de extrañar, pues, que la IAF esté resultando ser tan esquiva.

DIMENSIONES QUE FALTAN

Casi todos los sistemas generalistas actuales se centran en la cognición. Anderson, por ejemplo, pretende especificar "cómo se conectan entre sí todos los subcampos de la psicología cognitiva". (¿"Todos" los subcampos? Aunque aborda el control motor, no trata del tacto o de la propiocepción, que a veces aparecen en robótica). Una IA verdaderamente general abarcaría la motivación y la emoción.

Unos pocos científicos de IA lo han reconocido. Tanto Marvin Minsky como Sloman han escrito con gran agudeza sobre la arquitectura computacional de la mente en su conjunto, aunque ninguno haya construido un modelo.[32]

El modelo de ansiedad de Sloman, MINDER, se describe en el capítulo III. Su trabajo (y la teoría psicológica de Dietrich Dorner) le han servido de inspiración a Joscha Bach para MicroPsi: una IAF basada en siete "motivos" diferentes que usa disposiciones "emocionales" en la planificación y selección de acciones. También ha tenido influencia en el sistema LIDA mencionado anteriormente (véase el capítulo VI), pero incluso estos distan mucho de una verdadera IAF. El manifiesto visionario de IA de Minsky, "Pasos hacia la Inteligencia Artificial", identificaba tantos obstáculos como premisas.[33] Todavía hay que superar muchos de los primeros.

Como el capítulo III debería ayudar a demostrar, la IAF a nivel humano aún no se vislumbra. Muchos profesionales de la IA difieren. Algunos incluso añaden que la IAF se convertirá en breve en IAS ("s" de sobrehumano) y en consecuencia marginará al Homo sapiens: véase el capítulo VII.

III

LENGUAJE, CREATIVIDAD, EMOCIÓN

Algunos ámbitos de la IA parecen especialmente complejos, como el lenguaje, la creatividad y la emoción. Si la IA no es capaz de replicarlos, es ilusorio esperar que llegue la IAF (Inteligencia Artificial Fuerte).

En todos estos ámbitos, quintaesencia de lo "humano", se ha conseguido más de lo que muchos se imaginan. No obstante, persisten algunas dificultades significativas y han sido replicados solo hasta cierto punto. (De si los sistemas de IA podrán tener alguna vez entendimiento, creatividad o emoción reales se habla en el capítulo VI. Ahora, nuestra pregunta es si puede parecer que los tienen).

LENGUAJE

Innumerables aplicaciones de IA usan el procesamiento de lenguajes naturales (PLN o NLP por sus siglas en inglés). La mayoría se centra en la "comprensión" por parte del ordenador del lenguaje que se le presenta, no de su propia producción lingüística. Esto se debe a que generar PLN es todavía más difícil que aceptarlo.

Las dificultades atañen tanto al contenido temático como a las formas gramáticas. Por ejemplo, vimos en el capítulo II que se pueden utilizar secuencias de acciones conocidas ("scripts") como esbozo para relatos generados mediante IA, pero que la representación del conocimiento de base contenga motivos humanos suficientes como para que el relato sea interesante es otro asunto. Un sistema que se puede adquirir en el mercado escribe

sumarios anuales sobre las fluctuaciones de la situación económica de una empresa, pero genera “historias” muy aburridas. Las novelas y los culebrones generados por ordenador sí que existen, aunque no ganarían ningún premio a la sutileza. (Las traducciones / resúmenes realizados con ia de textos escritos por humanos seguramente sean mucho más ricas, pero es gracias a esos autores humanos).

En cuanto a la forma gramatical, la prosa automática es a veces gramaticalmente incorrecta y muy burda por lo general. Una narración creada por una IA sobre una partida de tres en raya puede contener estructuras causales / subclausales que se ajusten a la dinámica del juego de manera totalmente adecuada.[1] Pero las posibilidades y estrategias del tres en raya se comprenden de sobra. Describir la sucesión de pensamientos o acciones de los protagonistas de la mayoría de los relatos humanos con la misma elegancia sería mucho más complejo.

En cuanto a la aceptación del lenguaje, algunos sistemas de IA son simples hasta el aburrimiento: solo precisan reconocer palabras clave (como en los “menús” de las tiendas en línea) o predecir las palabras listadas en un diccionario (como el teclado predictivo al escribir mensajes de texto). Otros son muchísimo más sofisticados.

Unos pocos requieren reconocimiento de voz, ya sea de palabras sueltas (como en la compra por teléfono automatizada) o del habla continua (como el subtitulado en vivo de la TV y la escucha telefónica). En este último caso, el objetivo puede ser detectar palabras específicas (como bomba y yihad) o, lo que es más interesante, captar el sentido de la frase como un todo. Esto es procesamiento de lenguajes naturales (PLN) pero con todos los aditamentos: primero hay que distinguir las palabras mismas dichas por muchas voces diferentes y con diferentes acentos locales / extranjeros. (Las distinciones entre palabras vienen incluidas en los textos impresos). El aprendizaje profundo (véase el capítulo IV) ha hecho posibles avances significativos en el procesamiento del habla.[2]

Ejemplos impresionantes de lo que parece comprensión de una frase completa son la traducción automática, la minería de datos de grandes recopilaciones de textos de lenguaje natural, el resumen de artículos de periódicos y revistas y la respuesta de preguntas de amplio espectro (que cada vez se usa más en las búsquedas de Google y en la aplicación Siri para el iPhone).

Pero ¿pueden estos sistemas apreciar de verdad el lenguaje? ¿Pueden hacer frente a la gramática, por ejemplo?

En los primeros tiempos de la IA, se suponía que la comprensión del lenguaje precisaba de análisis sintáctico. Se realizaron esfuerzos considerables para escribir programas que lo llevaran a cabo. El ejemplo más destacado (que atrajo la atención de infinidad de personas que no habían oído hablar de la IA antes o que la habían descartado por imposible) fue SHRDLU, de Terry Winograd, escrito en el MIT (Instituto Tecnológico de Massachusetts) a principios de la década de 1970.[3]

Este programa aceptaba instrucciones en inglés para que un robot construyera estructuras hechas de bloques de colores y averiguaba cómo tenían que moverse ciertos bloques para alcanzar el objetivo. Fue muy influyente por muchas razones, algunas de las cuales se podían aplicar en la IA general. Lo relevante es su capacidad sin precedentes para asignarles estructura gramatical detallada a frases complejas, como: “¿Cuántos huevos habrías ido a utilizar para la tarta si no hubieses sabido que la receta de tu abuela estaba mal?” (¡Inténtelo!).

Para propósitos tecnológicos, SHRDLU resultó ser una decepción. El programa contenía muchos fallos, así que solo lo podían utilizar unos cuantos investigadores muy cualificados. En aquella época se construyeron otros cruncheadores de sintaxis, que tampoco se podían generalizar para textos del mundo real. En definitiva, pronto quedó claro que el análisis de sintaxis sofisticada es demasiado difícil para los sistemas que se pueden adquirir en el mercado.

La sintaxis sofisticada no era el único problema; en el uso de la lengua humana, el contexto y la relevancia también importan. No parecía obvio que la IA pudiese dominarlos nunca.

De hecho, el gobierno de Estados Unidos había declarado imposible la traducción automática en su informe ALPAC de 1964 (el acrónimo se refería al Automatic Language Processing Advisory Committee, comité asesor para el procesamiento del lenguaje natural).[4] Además de predecir que a muy pocos les interesaría hacerla comercialmente viable (aunque la ayuda automática para traductores humanos sí que podría ser factible), el informe sostenía que los ordenadores forcejearían con la sintaxis, serían derrotados por el contexto y que, sobre todo, serían ciegos ante la relevancia.

Fue una bomba para la traducción automática (cuya financiación se agotó prácticamente de un día para el otro) y para la IA en general. Se interpretó de manera generalizada como una demostración de la futilidad de la IA. En el éxito de ventas Computers and Common Sense [Ordenadores y sentido común] ya se había afirmado (en 1961) que la IA era malgastar el dinero de los contribuyentes.[5] Ahora, parecía que los grandes expertos gubernamentales concordaban con eso. En consecuencia, dos universidades estadounidenses que iban a abrir departamentos de IA cancelaron sus planes.

El trabajo sobre IA siguió de todas maneras y cuando SHRDLU, experto en sintaxis, apareció en escena unos años después, pareció ser una vindicación triunfal de la inteligencia artificial simbólica. Pero no tardaron en surgir las dudas. Por eso, el procesamiento de lenguajes naturales (PLN) recurrió cada vez más al contexto en vez de a la sintaxis.

Unos cuantos investigadores se habían tomado en serio el contexto semántico ya a principios de la década de 1950. El grupo de Margaret Masterman en Cambridge, Inglaterra, había abordado la traducción automática (y la búsqueda y recuperación de

información) usando un tesauro en vez de un diccionario. Veían la sintaxis como “esa parte del lenguaje muy superficial y redundante que se deja de lado cuando se tiene prisa”[6] y se centraron en grupos de palabras más que en palabras sueltas. En vez de intentar traducir palabra por palabra, buscaron en el texto circundante palabras de significado similar. Gracias a esto (cuando funcionaba) se podían traducir correctamente palabras ambiguas. Así banco podía traducirse (en francés) como rive o banque, dependiendo de si en el contexto había palabras como agua o dinero respectivamente.

Este enfoque contextual basado en el tesauro podía reforzarse examinando también las palabras que solían concurrir a pesar de tener significados diferentes (como pez y agua). Y esto, con el paso del tiempo, es lo que ocurrió. Además de distinguir varios tipos de similitudes léxicas –sinónimos (vacío / vacante), antónimos (vacío / lleno), pertenencia a una clase (pez / animal) e inclusión (animal / pez), nivel de clase compartido (bacalao / salmón) y parte / todo (aleta / pez)–, la traducción automática actual también reconoce la concurrencia temática (pez / agua, pez / banco, pescado / patatas, etcétera).

Ahora queda claro que no es necesario dominar una sintaxis sofisticada para resumir, preguntar o traducir un texto de lenguaje natural. El procesamiento de lenguajes naturales actual depende más de la fuerza (potencia computacional) que de la maña (análisis gramatical). Las matemáticas –específicamente, la estadística– han superado a la lógica y el aprendizaje automático (que incluye el aprendizaje profundo, pero no solo) ha desplazado al análisis sintáctico. Estos nuevos enfoques de PLN, que abarcan desde textos escritos al reconocimiento de habla, son tan eficientes que se toma como norma de aceptación de aplicaciones prácticas una tasa del 95% de éxito.

En el procesamiento de lenguajes naturales moderno, unos ordenadores potentes hacen búsquedas estadísticas de recopilaciones enormes (“corpus”) de textos (para la traducción automática, son pares de traducciones realizadas por humanos)

para encontrar patrones de palabras tanto comunes como inesperados. Pueden reconocer la probabilidad estadística de pez / agua o pez / renacuajo o pescado con patatas / sal y vinagre. Y (como se señaló en el capítulo II) el PLN ahora es capaz de aprender a crear “vectores de palabras” que representan las nubes de significado probabilístico que participan en un concepto dado.[7] En general, sin embargo, se hace hincapié en las palabras y en las frases, no en la sintaxis. No se ignora la gramática: se pueden asignar etiquetas como adjetivo y adverbio a algunas palabras de los textos que se estén examinando,[8] ya sea automáticamente o a mano, pero el análisis sintáctico se usa poco.

Ni siquiera destaca mucho el análisis semántico detallado. La semántica “composicional” usa la sintaxis para analizar los significados de los enunciados, pero en laboratorios de investigación, no en aplicaciones a gran escala. El razonador de “sentido común” CYC contiene representaciones semánticas de sus conceptos (palabras) relativamente completas y, en consecuencia, las “entiende” mejor (véase el capítulo II). Pero esto sigue siendo inusual.

La traducción automática actual puede ser asombrosamente correcta. Algunos sistemas se restringen a un conjunto pequeño de temas, pero otros son más abiertos. Google Translate ofrece traducción automática sobre temas sin restricciones a más de doscientos millones de usuarios cada día. La Unión Europea (para 24 idiomas), la OTAN, Xerox y General Motors usan SYSTRAN a diario.

Muchas de estas traducciones, incluyendo los documentos de la UE, son casi perfectas (porque en los textos originales solo se utiliza un conjunto limitado de palabras). Muchas otras son imperfectas, pero aun así casi inteligibles, porque los lectores informados ignoran los errores gramaticales y las elecciones poco gráciles de palabras, como cuando se escucha a un hablante no nativo. Algunas requieren una edición posterior mínima. (Para el japonés puede que se necesite edición previa y posterior. El japonés no tiene palabras

segmentadas, como el pretérito en inglés work-ed, y el orden de las frases está invertido. La equiparación automática de lenguas de diferentes grupos lingüísticos suele ser difícil).

En suma, los resultados de la traducción automática son por lo general lo suficientemente buenos para que el usuario humano los comprenda. De la misma forma, los programas de procesamiento de lenguajes naturales (PLN) monolingües que resumen periódicos suelen dejar claro si el artículo merece ser leído en su totalidad. (De todos modos, la traducción perfecta seguramente sea imposible. Por ejemplo, pedir una manzana en japonés requiere un lenguaje que refleje la diferencia de estatus social, cuando en inglés no existen distinciones equivalentes).

La traducción en tiempo real disponible en aplicaciones de IA como Skype es menos eficaz. Esto se debe a que el sistema tiene que reconocer habla, no texto escrito, en el que las palabras individuales están separadas con claridad.

Otras dos aplicaciones destacadas de procesamiento de lenguajes naturales son modalidades de extracción de información: la búsqueda ponderada (iniciada por el grupo de Masterman en 1976) y la minería de datos. El motor de búsqueda de Google, por ejemplo, busca términos ponderados por su relevancia, que se determina mediante estadística, no semántica (esto es, sin comprenderlos). La minería de datos es capaz de encontrar patrones de palabras insospechados por el usuario humano. Utilizada desde hace tiempo en investigaciones de mercado para productos y marcas, se está aplicando ahora (en muchos casos mediante aprendizaje profundo) a “big data”: recopilaciones enormes de textos (a veces multilingües) o imágenes, como informes científicos, registros médicos o entradas de redes sociales e internet.

Entre las aplicaciones de minería de big data se incluyen la vigilancia y el contraespionaje y el control de la opinión pública por parte de gobiernos, responsables políticos y científicos de

disciplinas sociales. Estas consultas comparan las opiniones de distintos subgrupos: hombres / mujeres, jóvenes / mayores, norte / sur y así. Por ejemplo, Demos, un equipo de expertos del Reino Unido (junto a un equipo de análisis de datos de PLN de la universidad de Sussex) ha analizado muchos miles de mensajes de Twitter relativos a misoginia, grupos étnicos y policía. Se puede buscar en los estallidos repentinos de tuits después de sucesos específicos (“tuitcidentes”) para descubrir, por ejemplo, cambios en la opinión pública sobre la reacción de la policía ante un suceso particular.

Queda por verse si el PLN en big data producirá resultados útiles de manera fiable. Muchas veces, la minería de datos (mediante el “análisis de sentimiento”) pretende medir, además del nivel del interés público, su valoración. Pero esto no es tan sencillo. Por ejemplo, un tuit que contenga un epíteto racial aparentemente despectivo (y que por eso se codifica automáticamente como sentimiento “negativo”), puede que de hecho no sea peyorativo. Un juez humano, al leerlo, puede ver que el término se está usando (en este caso) como una marca positiva de identidad grupal o como una descripción neutral (por ejemplo, “el mercado paqui de la esquina”), no como insulto. (La investigación de Demos descubrió que en solo una pequeña proporción de tuits los términos raciales / étnicos eran agresivos en realidad).[9]

En casos así, el juicio humano confiará en el contexto, por ejemplo, las otras palabras incluidas en el tuit. Puede ser posible ajustar los criterios de búsqueda de la máquina para que haga menos atribuciones de “sentimiento negativo”. Aunque… puede ser que no. Tales juicios suelen ser polémicos. Incluso cuando se aceptan, puede resultar difícil identificar aquellos aspectos del contexto que justifiquen la interpretación humana.

Es un solo ejemplo de la dificultad de localizar la relevancia en términos computacionales (o incluso verbales).

Puede parecer a primera vista que dos aplicaciones de procesamiento de lenguajes naturales (PLN) muy conocidas contradicen esa afirmación: Siri de Apple y WATSON de IBM.

Siri es un asistente personal (basado en reglas), un “bot conversacional” que habla y puede responder rápidamente muchas preguntas diferentes. Tiene acceso a todo internet, incluyendo Google Maps, Wikipedia, el siempre actualizado The New York Times y listas de servicios locales como taxis y restaurantes. También recurre al potente buscador de respuestas Wolfram Alpha, capaz de utilizar el razonamiento lógico para averiguar (no solo encontrar) la respuesta a un amplio rango de preguntas fácticas.

Siri acepta una pregunta oral del usuario (a cuya voz y dialecto se va adaptando gradualmente) y la responde usando la búsqueda web y el análisis conversacional. El análisis conversacional estudia cómo se organiza la secuencia de temas en una conversación, cómo se ordenan las interacciones tales como explicación y acuerdo y le permite a Siri examinar preguntas como ¿Qué quiere el interlocutor? y ¿Cómo debería responder? y (hasta cierto punto) adaptarse a los intereses y preferencias del usuario individual.

En suma, Siri parece ser sensible no solo a la relevancia temática, sino también a la relevancia personal, así que en apariencia es impresionante. Sin embargo, es fácil manipularla para que ofrezca respuestas ridículas y, si el usuario se desvía del ámbito de los hechos, Siri se pierde.

WATSON, también, se concentra en los hechos. Como recurso comercial (con 2.880 procesadores de núcleo) para manejar big data ya se usa en algunos centros de atención telefónica y se está adaptando para aplicaciones médicas como evaluador de terapias para el cáncer. Pero no se limita a responder preguntas directas como Siri, también puede resolver los acertijos que aparecen en el juego de cultura general Jeopardy![10].

En Jeopardy! no se hacen preguntas directas a los jugadores, se les da una pista y tienen que averiguar la pregunta a la que corresponde. Por ejemplo, “El 9 de mayo de 1921, esta aerolínea de siglas perfectas abrió su primera oficina de atención al viajero en Ámsterdam” y deberían responder “¿Qué es KLM?”.

WATSON puede hacer frente a ese desafío y a muchos otros. A diferencia de Siri, su versión para jugar a Jeopardy! no tiene acceso a internet (aunque la versión médica sí) y ninguna noción de estructura conversacional. Tampoco puede descubrir respuestas mediante razonamiento lógico. En cambio, realiza una búsqueda estadística masivamente paralela en una base de datos enorme, aunque cerrada, que contiene documentos (innumerables revistas y libros de referencia, más The New York Times) que lo proveen de datos desde la lepra a Liszt, del hidrógeno a Hidra y etcétera. Cuando juega a Jeopardy!, cientos de algoritmos creados especialmente para reflejar las probabilidades inherentes del juego guían su búsqueda. Y es capaz de aprender de las respuestas de sus rivales humanos.

En 2011, WATSON rivalizó con el momento Kasparov de su primo de IBM Deep Blue (véase el capítulo II), venciendo aparentemente a los dos campeones humanos. (“Aparentemente” porque el ordenador reacciona instantáneamente, mientras que los seres humanos necesitan un tiempo de reacción antes de presionar el botón). Pero, igual que Deep Blue, no siempre gana.

En una ocasión perdió porque, aunque se fijó correctamente en la pierna de un atleta particular, no se percató de que el hecho crucial de los datos almacenados era que al atleta le faltaba una pierna. Ese error no volverá a cometerlo, porque los programadores de WATSON han señalado la importancia de la palabra “faltante”, pero cometerá otros. Incluso en contextos de búsqueda de información prosaica, la gente suele confiar en juicios de relevancia que quedan fuera del alcance de WATSON. Por ejemplo, una prueba preguntaba por la identidad de los dos discípulos de Jesús cuyos nombres terminan con la misma letra y están entre los diez

nombres de bebés más frecuentes. La respuesta era "Tomás y Andrés", que WATSON encontró de inmediato. El campeón humano también supo la respuesta, pero su primera idea había sido "Tomás y Judas". Recordó haberla rechazado porque pensó "Por alguna razón, no creo que Judas sea un nombre que se le ponga mucho a los bebés". WATSON no podría haber hecho eso.

Los juicios humanos de relevancia suelen ser aún menos obvios que ese y mucho más sutiles para el procesamiento de lenguajes naturales actual. De hecho, la relevancia es una versión lingüístico / conceptual del implacable "problema del marco" de la robótica (véase el capítulo II). Muchos argumentarían que un sistema no humano nunca podrá dominar la relevancia por completo. Si esto se debe solo a la complejidad considerable que entraña o al hecho de que la relevancia está arraigada en nuestra forma de vida específicamente humana, se trata en el capítulo VI.

CREATIVIDAD

La creatividad (la capacidad de producir ideas o artefactos que sean nuevos, sorprendentes y valiosos) es el cénit de la inteligencia humana y es necesaria para una IAF a nivel humano. Pero se la suele considerar misteriosa. No está claro cómo se les ocurren ideas nuevas a las personas, y ya no digamos a los ordenadores.

Ni siquiera reconocer la creatividad es tan sencillo: suele haber desacuerdo sobre si una idea es creativa. Algunas controversias tratan de si y en qué sentido es de verdad una idea nueva: puede serlo solo para el individuo en cuestión o para toda la historia humana (lo que ilustra la creatividad "individual" e "histórica" respectivamente). En ambos casos, puede ser más o menos similar a las ideas precedentes, dejando margen para controversias posteriores. Otros conflictos giran alrededor de la valoración (que requiere de consciencia funcional y a veces también de consciencia fenoménica: véase el capítulo VI). Un grupo social puede valorar una idea y otros grupos no. (Piense el lector en el desprecio de los jóvenes de hoy hacia alguien que atesore discos de Abba).

Se asume por lo general que la IA podría no tener nada interesante que decir sobre la creatividad, pero la tecnología que utiliza ha generado muchas ideas históricamente nuevas, sorprendentes y valiosas que aparecen, por ejemplo, en el diseño de motores, productos farmacéuticos y varios tipos de arte computacional.

Además, los conceptos de IA ayudan a explicar la creatividad humana y nos permiten distinguir tres tipos: combinatoria, exploratoria y transformacional,[11] que comportan diferentes mecanismos psicológicos y suscitan diferentes tipos de asombro.

En la creatividad combinatoria, las ideas conocidas se combinan de maneras desconocidas. Algunos ejemplos son el collage visual, las imágenes poéticas y las analogías científicas (el corazón es una bomba, el átomo es un sistema solar). La nueva combinación aporta una sorpresa estadística: era algo improbable, como que un profano gane el Derby. Pero es inteligible, luego valiosa. Cuán valiosa depende de los juicios de relevancia de los que se ha hablado antes.

La creatividad exploratoria es menos personalista, ya que se apoya en alguna manera de pensar que ya tiene valor cultural (por ejemplo, estilos pictóricos o musicales o subcampos de la química o las matemáticas). Se usan reglas estilísticas (en gran medida de manera inconsciente) para producir la nueva idea, tal como la gramática genera nuevas frases. El artista / científico puede explorar el potencial del estilo de forma incuestionable. O puede forzarlo y ponerlo a prueba, descubriendo lo que puede y no puede crear. Puede incluso modificarlo, alterando levemente (por ejemplo, debilitando / reforzando) una regla. La nueva estructura, a pesar de su novedad, será reconocida como parte de una familia estilística aceptada.

La creatividad transformacional es sucesora de la creatividad exploratoria, que por lo general se desencadenada debido a la

frustración que causan los límites del estilo existente. Gracias a ella, se pueden modificar una o más constricciones estilísticas de manera radical (mediante el abandono, la negación, la complementación, la sustitución, la adición…): así se crean estructuras nuevas que no podrían haberse creado antes. Estas nuevas ideas son profundamente asombrosas, porque aparentan ser imposibles. Al principio suelen ser ininteligibles, ya que no se pueden entender por completo en los términos del modo de pensar que antes se aceptaba. Sin embargo, deben ser inteligiblemente cercanas al modo de pensar previo para que se las acepte. (A veces, este reconocimiento tarda años).

Los tres tipos de creatividad se dan en la IA, muchas veces con resultados que los observadores atribuyen a seres humanos (en efecto, pasar la prueba de Turing: véase el capítulo VI), pero no en la proporción que cabría esperar.

En concreto, hay muy pocos sistemas combinatorios. Podría pensarse que es fácil copiar la creatividad combinatoria; al fin y al cabo, qué hay más simple que hacer que un ordenador produzca asociaciones desconocidas de historias que ya tiene almacenadas. Los resultados serán por lo general históricamente novedosos y (estadísticamente) sorprendentes, pero para que también sean valiosos deben ser mutuamente pertinentes. No es sencillo, como hemos visto. Los programas creadores de chistes mencionados en el capítulo II usan modelos de chistes para ayudarse a ser relevantes. De la misma forma, el razonamiento basado en casos de la IA simbólica construye analogías gracias a las similitudes estructurales pre-codificadas.[12] Así, su creatividad “combinatoria” contiene también una fuerte dosis de creatividad exploratoria.[13]

Por otro lado, cabría esperarse que la IA no pudiese imitar nunca la creatividad transformacional. Esta expectativa, también, es errónea. Cierto es que cualquier programa puede hacer solamente lo que es capaz de hacer en potencia, pero los programas evolutivos pueden transformarse a sí mismos (véase el capítulo V) y evaluar sus ideas recién transformadas, pero solo si el programador ha

proporcionado criterios claros para la selección. Estos programas se utilizan de manera rutinaria para buscar aplicaciones innovadoras de IA, como diseñar nuevos instrumentos científicos o medicamentos.

Sin embargo, no es un camino mágico que conduzca a la IAF. Rara vez se puede garantizar que haya resultados valiosos. Algunos programas evolutivos (de matemáticas o ciencias) pueden encontrar la solución óptima de manera fiable, pero hay muchos problemas que no se pueden definir mediante la optimización. La creatividad transformacional es arriesgada, porque se rompen las reglas aceptadas previamente. Debe evaluarse cualquier estructura nueva o sobreviene el caos. Pero las funciones de adecuación de la IA vigentes las definen los humanos: los programas no pueden adaptarse / evolucionar de manera independiente.

La creatividad exploratoria es el tipo de creatividad que mejor casa con la IA. Hay ejemplos infinitos. A algunas novedades exploratorias de IA en ingeniería (incluyendo una generada por un programa del diseñador del CYC: véase el capítulo II) se les han concedido patentes. Aunque una idea patentada pueda no parecerle "evidente a un experto en la materia", es posible que, inesperadamente, pertenezca al estilo que se está estudiando. Unos cuantos ejercicios de IA son indistinguibles de algunos logros humanos destacados, como la composición musical al estilo de Chopin o Bach.[14] (¿Cuántos seres humanos pueden hacer eso?).

Sin embargo, la IA exploratoria depende en lo crucial del juicio humano, ya que alguien debe reconocer (y establecer claramente) las reglas estilísticas pertinentes, lo que suele ser difícil. Un experto mundial en las Casas de la Pradera de Frank Lloyd Wright abandonó el intento de describir su estilo arquitectónico y lo declaró "ocultista". Tiempo después, una "gramática geométrica" computable generó de manera indefinida muchos diseños de Casas de la Pradera, incluidas las cuarenta originales y ninguna inverosímil.[15] Pero, en última instancia, el analista humano era responsable del éxito del sistema. Solo si una IAF pudiese analizar estilos (de arte o científicos) por sí misma, serían "obra suya" sus intentos. A pesar de

que han sido reconocidos algunos ejemplos recientes (muy limitados) de estilos artísticos realizados por el aprendizaje profundo (véanse los capítulos II y IV), es mucho pedir.

La IA ha permitido que los artistas humanos desarrollen una nueva forma artística: el arte generado por ordenador, que abarca la arquitectura, las artes gráficas, la música, la coreografía y la literatura (esta con menos éxito, dadas las dificultades que tiene el procesamiento de lenguajes naturales con la sintaxis y la relevancia). En el arte digital, el ordenador no es una simple herramienta comparable a un nuevo pincel que ayuda al artista a hacer cosas que podría haber hecho de todas formas. Más bien, la obra no podría haberse realizado, quizá ni tan siquiera imaginado, sin él.[16]

El arte generado por ordenador ejemplifica los tres tipos de creatividad. Por las razones dadas anteriormente, casi ningún arte generado por ordenador es combinatorio. (The Painting Fool de Simon Colton ha producido collages visuales relacionados con la guerra, pero había recibido la instrucción específica de buscar imágenes asociadas con “guerra” que ya estaban disponibles en su base de datos).[17] La mayoría es exploratorio o transformacional.

A veces, el ordenador genera la obra artística de manera totalmente independiente ejecutando el programa escrito por el artista. Así, el AARON de Harold Cohen produce dibujos lineales e imágenes coloreadas sin ayuda (a veces crea colores tan audaces y hermosos que Cohen dice que es mejor colorista que él mismo).[18]

En el arte interactivo, por el contrario, la forma de la obra artística final depende en parte de la contribución del público, que puede o no tener control deliberado sobre lo que ocurre. Algunos artistas interactivos consideran al público como compañero de creación, otros como un mero factor causal que afecta sin saber a la obra artística de varias formas (y otros, como Ernest Edmonds, usan ambos criterios). En el arte evolutivo, del que William Latham y Jon McCormack[19] son un ejemplo, los resultados los genera el

ordenador de forma continua, pero la selección la suelen hacer el artista o el público.

En suma, la creatividad de la IA tiene muchas aplicaciones que a veces igualan, o incluso exceden, a los modelos humanos en algún pequeño rincón de la ciencia o del arte, pero igualar la creatividad humana en general es un asunto muy diferente. La IAF queda tan lejos como siempre.

IA Y EMOCIÓN

La emoción, como la creatividad, se suele considerar algo completamente ajeno a la IA. Además de que se intuye que no es plausible, el hecho de que los estados de ánimo y las emociones dependan de neuromoduladores que se difunden por el cerebro parece descartar los modelos de la emoción de la IA.

Durante muchos años, hasta los mismos científicos especializados en IA parecían estar de acuerdo. Con unas pocas excepciones en las décadas de 1960 y 1970 (en concreto, Herbert Simon,[20] quien consideraba que la emoción estaba involucrada en el control cognitivo, y Kenneth Colby,[21] que creó modelos de neurosis y paranoia interesantes, aunque demasiado ambiciosos) ignoraron la emoción.

Hoy, las cosas son diferentes: se ha simulado la neuromodulación (en GasNets: véase el capítulo IV). Además, muchos grupos de investigación sobre IA están empezando a abordar la emoción. La mayor parte de estas investigaciones (no todas) es teóricamente superficial y potencialmente lucrativa, ya que se orienta al desarrollo de “robots acompañantes”.[22]

Son sistemas de IA (algunos con una pantalla como base, otros robots ambulantes) diseñados para interactuar con personas de manera que (además de ser útiles en la práctica) le generan confortación afectiva, e incluso placer, al usuario. La mayoría se destina a ancianos y/o discapacitados y a personas con demencia

senil incipiente. Algunos están concebidos para bebés o niños pequeños, otros son “juguetes para adultos” interactivos. En suma: cuidadores computerizados, niñeras robot y compañeros sexuales.

Las interacciones entre humanos y ordenadores en cuestión incluyen: ofrecer recordatorios sobre compras, medicación y visitas familiares; hablar y ayudar a compilar un diario personal; recordar los horarios y conversar sobre programas de TV, incluyendo las noticias; preparar / traer comida y bebida; vigilar los signos vitales (y el llanto de los bebés); y hablar y moverse de manera sexualmente incitante.

Muchas de estas tareas comportarán emociones por parte de la persona. En cuanto al compañero de IA, puede que sea capaz de reconocer emociones en el usuario humano y/o que pueda reaccionar de formas que parezcan emocionales. Por ejemplo, cuando el usuario sienta tristeza (provocada, quizá, por la mención de la pérdida de un ser querido), podría suscitar alguna muestra de compasión por parte de la máquina.

Los sistemas de IA ya pueden reconocer emociones humanas de varias formas, por ejemplo, fisiológicas: observando la frecuencia respiratoria y la respuesta galvánica de la piel de la persona; verbales: observando la velocidad de habla y la entonación, así como el vocabulario; y visuales: analizando las expresiones del rostro. Por el momento, todos estos métodos son relativamente rudimentarios. Las emociones del usuario se pueden pasar por alto o malinterpretarse fácilmente.

El desempeño emocional del acompañante computerizado suele ser verbal y se basa en el vocabulario (y la entonación, si el sistema puede generar habla). Pero, aunque el sistema esté atento a captar las palabras clave conocidas que utiliza el usuario, responde de maneras muy estereotipadas. En ocasiones, puede citar a una figura importante humana o un poema relacionado con algo que ha dicho el usuario, quizá en su diario, pero las dificultades del procesamiento de lenguajes naturales (PLN) implican que no es

probable que el texto generado por ordenador sea apropiado ni sutil. Puede que incluso sea inaceptable: un compañero incapaz de ofrecer siquiera la apariencia de una compañía verdadera puede irritar y frustrar al usuario. De la misma forma, un robot gato que ronronee puede molestar al usuario en vez de transmitir tranquilamente un relajado contento.

O tal vez no: Paro, un mimoso “bebé foca” interactivo con encantadores ojos negros y pestañas exuberantes, parece ser beneficioso para muchos ancianos y/o personas con demencia senil. (Las versiones futuras controlarán los signos vitales y alertarán a los cuidadores humanos de la persona en caso necesario).

Algunas IA acompañantes pueden usar sus propias expresiones faciales y su mirada para reaccionar de manera aparentemente emocional. Unos cuantos robots poseen “piel” flexible, recubierta con un simulacro de musculatura facial humana, cuya configuración puede sugerir (al observador humano) hasta una docena de emociones. Los sistemas con pantalla suelen mostrar el rostro de un personaje virtual cuyas expresiones cambian según las emociones que se supone que siente (¿él / ella?). Sin embargo, todas estas cosas [sic] corren el riesgo de caer en el así llamado “valle inquietante”: la gente se siente incómoda o incluso profundamente turbada al encontrarse con criaturas muy similares a los seres humanos, pero no lo bastante similares. Los robots o los avatares de la pantalla con rostros no exactamente humanos pueden, por lo tanto, percibirse como amenazantes.

Es cuestionable la ética de ofrecer esta pseudocompañía a gente emocionalmente necesitada (véase el capítulo VII). Es cierto que algunos sistemas interactivos entre humanos y ordenadores (por ejemplo, Paro) parecen proporcionar deleite y hasta un bienestar duradero a personas cuyas vidas, si no, parecerían vacías; pero ¿basta con eso?

Hay escasa profundidad teórica en los modelos “de compañía”. Los aspectos emocionales de las IA acompañantes están en fase de

desarrollo y tienen propósitos comerciales. No hay intención de que utilicen sus emociones para resolver sus propios problemas, ni para esclarecer el papel que juegan las emociones en el funcionamiento de la mente en su conjunto. Es como si estos investigadores de IA viesen las emociones como extras opcionales que se pueden ignorar a no ser que sean inevitables en algún contexto desordenadamente humano.

Esa actitud despectiva estaba muy extendida en la IA hasta hace relativamente poco. Ni siquiera en el trabajo de Rosalind Picard sobre “computación afectiva”, que rehabilitó las emociones a finales de la década de 1990, se las analizó en profundidad.[23]

Una razón por la que la IA ignoró la emoción (y las perspicaces observaciones de Simon sobre ella) durante tanto tiempo es que la mayoría de los psicólogos y filósofos también lo hacían. Dicho de otro modo, no pensaron que la inteligencia fuese algo que requiriese emoción. Al contrario, se asumía que el afecto perturbaba la resolución de problemas y la racionalidad. La idea de que la emoción puede ayudar a decidir qué hacer y la mejor manera de hacerlo no estaba de moda.

Al final fue cobrando mayor importancia, gracias en parte a los adelantos en psicología clínica y neurociencia. Su entrada en la IA se debió también a dos científicos, Marvin Minsky y Aaron Sloman,[24] que llevaban mucho tiempo estudiando la mente en su conjunto, en vez de restringirse –como la mayoría de sus colegas– a un rinconcito de la mentalidad.

Por ejemplo, el proyecto permanente de Sloman CogAff se centra en el papel de la emoción en la arquitectura computacional de la mente. CogAff ha influido en el modelo de consciencia LIDA, lanzado en 2011 y que sigue ampliándose (véase el capítulo VI). También ha inspirado el programa MINDER, iniciado por el grupo de Sloman a finales de la década de 1990.

MINDER simula (los aspectos funcionales de) la ansiedad que experimenta una enfermera a la que han dejado sola para cuidar de varios bebés. Tiene solo unas cuantas tareas: alimentarlos, tratar de evitar que caigan en zanjas y llevarlos a un puesto de primeros auxilios en caso de que suceda. Y tiene solo unos cuantos motivos (objetivos): alimentar al bebé; colocar al bebé detrás de una valla protectora, en caso de que ya haya una; sacar al bebé de la zanja para llevarlo a primeros auxilios; patrullar la zanja; construir una valla; trasladar al bebé a una distancia segura de la zanja; y, si no hay otro motivo activado, dar vueltas por la guardería.

Así que es muchísimo más simple que una enfermera real (aunque más complejo que un programa de planificación típico, que tiene un solo objetivo final). No obstante, es proclive a perturbaciones emocionales comparables a varios tipos de ansiedad.

La enfermera simulada tiene que reaccionar de manera apropiada a las señales visuales del medio. Algunas de estas señales desencadenan (o inspiran) objetivos que son más urgentes que otros: un bebé que gatee hacia la zanja necesita atención antes que un bebé que solo está hambriento, y uno que está a punto de caer dentro de la zanja, todavía más.

Pero incluso estos objetivos que pueden diferirse quizá tengan que abordarse ulteriormente y su grado de urgencia puede aumentar con el tiempo. Así, se puede devolver a un bebé con hambre a su cuna si otro bebé está cerca de la zanja, pero habría que alimentar al bebé que lleve más tiempo esperando que le den de comer antes que a los que han comido los últimos.

En suma, la enfermera puede interrumpir sus tareas y abandonarlas o dejarlas para más tarde. MINDER tiene que decidir cuáles son las prioridades en cada momento. Debe tomar estas decisiones a lo largo de la sesión y estas pueden desembocar en repetidos cambios de comportamiento. Prácticamente ninguna tarea se puede completar sin interrupción, porque el medio (los bebés) le

impone al sistema muchísimas exigencias conflictivas que cambian constantemente. Igual que le ocurre a una enfermera real, la ansiedad aumenta y el desempeño empeora al aumentar el número de bebés, cada uno de los cuales es un agente autónomo impredecible. Sin embargo, la ansiedad es útil, ya que permite que la enfermera consiga alimentar a los bebés, aunque no con tranquilidad: la calma y la ansiedad son polos opuestos.

MINDER muestra varias formas en que las emociones pueden controlar el comportamiento, programando motivaciones competitivas de forma inteligente. Una enfermera humana experimenta [sic] varios tipos de ansiedad cada vez que cambia la situación. Pero la cuestión, en este caso, es que las emociones no son meros sentimientos, requieren de consciencia funcional y fenoménica (véase el capítulo VI). En concreto, son mecanismos computacionales que nos permiten programar motivaciones competitivas y sin los que no podríamos funcionar. (Así que el imperturbable señor Spock de Star Trek es una imposibilidad evolutiva). Para que podamos conseguir la IAF, emociones como la ansiedad tendrán que incluirse y utilizarse.

IV

REDES NEURONALES ARTIFICIALES

Las redes neuronales artificiales (RNA) están compuestas por muchas unidades interconectadas, cada una capaz de realizar una sola operación. Así descritas pueden sonar aburridas, pero pueden llegar a parecer casi mágicas. Desde luego han hechizado a los periodistas. Los “perceptrones” de Frank Rosenblatt,[1] máquinas fotoeléctricas que aprendieron a reconocer letras sin que se les enseñara de forma explícita, fueron promocionados con entusiasmo en los periódicos de la década de 1960. Las RNA dieron la gran campanada a mediados de los 80 y siguen siendo elogiadas cada poco en los medios. El bombo publicitario más reciente relacionado con las RNA concierne al aprendizaje profundo.

Las RNA tienen miríadas de aplicaciones, desde especular en bolsa y observar la fluctuación de las divisas a reconocer habla o caras. Pero lo que es intrigante es la forma en que funcionan.

Un pequeño grupo se ejecuta específicamente en hardware paralelo o quizá en una mezcla de hardware / wetware en la que se combinan neuronas reales con circuitos de silicio. Por lo general, sin embargo, la red se simula en una máquina de Von Neumann. Esto es, las RNA son máquinas virtuales de computación paralela implementadas en ordenadores clásicos (véase el capítulo I).

Su carácter fascinante viene, entre otras cosas, de que son muy diferentes a las máquinas virtuales de la IA simbólica. Las instrucciones secuenciales se sustituyen por procesamiento paralelo masivo, control desde arriba (top-down) mediante procesamiento desde abajo (bottom-up) y lógica probabilística. Y el aspecto

dinámico y constantemente cambiante de las RNA contrasta notablemente con los programas simbólicos.

Además, muchas redes tienen la propiedad asombrosa de autoorganizarse a partir de un comienzo aleatorio. (Los perceptrones de la década de 1960 también la tenían, de ahí su notoriedad en prensa). El sistema empieza con una arquitectura aleatoria (pesos de las unidades neuronales aleatorios y conexiones aleatorias) y se va adaptando a sí mismo de manera gradual para realizar la tarea requerida.

Las redes neuronales tienen muchas virtudes y han añadido capacidad de computación significativa a la IA. No obstante, también tienen deficiencias, como que no pueden producir esa IA verdaderamente general que se imaginó en el capítulo II. Por ejemplo, aunque algunas RNA puedan realizar inferencias o razonamientos aproximados, no son capaces de representar la precisión tan bien como la IA simbólica. “P: ¿Cuánto son 2 + 2? R: Muy probablemente 4”. ¿En serio? La jerarquía, también, es más difícil de modelar en las RNA. Algunas redes (recurrentes) pueden usar redes que interaccionan entre sí para representar jerarquías, pero solo hasta cierto punto.

Gracias al entusiasmo actual por el aprendizaje profundo, las redes de redes no son tan infrecuentes como solían. No obstante, siguen siendo relativamente sencillas. El cerebro humano debe contener redes innumerables en muchos niveles diferentes que interactuarán de maneras complejísimas. En suma, la IAF todavía queda muy lejos.

IMPLICACIONES MAYORES DE LAS RNA

Las RNA son un triunfo de la IA considerada como ciencia informática, pero sus implicaciones teóricas van mucho más allá. Debido a algunas similitudes generales con la memoria y los conceptos humanos, las RNA son interesantes para los neurocientíficos, los psicólogos y los filósofos.

El interés neurocientífico no es nuevo. De hecho, lo que Rosenblatt pretendía con aquellos perceptrones pioneros era que fuesen una teoría neuropsicológica y una fuente de gadgets útiles en la práctica. Las redes actuales, a pesar de sus muchas diferencias con el cerebro, son importantes en la neurociencia computacional.[2]

A los psicólogos también les interesan las RNA y los filósofos no les han ido a la zaga.[3] Por caso, un ejemplo de mediados de la década de 1980 causó furor fuera de las filas de la IA profesional.[4] Apareció una red que parecía aprender el uso del pretérito de forma muy parecida a los niños, sin cometer errores al principio, pero usando luego la sobrerregularización (así, volver / vuelto da lugar a volver / volvido) antes de conseguir construir correctamente los verbos tanto regulares como irregulares. Esto era posible porque las entradas que se le proporcionaban a la red imitaban las probabilidades variables de palabras que oye un niño normalmente: la red no aplicaba las reglas gramaticales innatas.

Esto era importante porque la mayoría de los psicólogos (y muchos filósofos) de la época habían aceptado las aseveraciones de Noam Chomsky de que los niños deben recurrir a reglas lingüísticas innatas para aprender gramática y que la sobrerregulación infantil era prueba irrefutable de que utilizaban esas reglas. La red del verbo en pretérito demostró que ninguna de esas aseveraciones era cierta. (No demostró, por supuesto, que los niños no tengan reglas innatas, solo que es innecesario que las tengan).

Otro ejemplo que despertó un amplio interés, inspirado originalmente en la psicología del desarrollo, es la investigación sobre “trayectorias representacionales”.[5] En este caso (como también en el aprendizaje profundo), los posibles datos iniciales confusos se recodifican en los niveles sucesivos para que se recojan las regularidades menos obvias junto a las más notables. Esto no se refiere solo al desarrollo infantil, sino también a los

debates psicológicos y filosóficos sobre lenguaje inductivo, ya que demuestra que se necesitan previsiones previas (la estructura computacional) para reconocer los patrones en los datos de entrada, y que hay limitaciones inevitables en el orden en que se aprenden diferentes pautas.

En suma, esta metodología de IA es teóricamente interesante en muchos sentidos y también tiene gran importancia comercial.

PROCESAMIENTO DISTRIBUIDO EN PARALELO

Una categoría en particular de las RNA llama mucho la atención: las que hacen procesamiento distribuido en paralelo (PDP).[6] De hecho, cuando la gente se refiere a las "redes neuronales" o al "conexionismo" (un término mucho menos usado en la actualidad), suelen referirse al PDP.

Debido a la manera en que funcionan, las redes de PDP comparten cuatro virtudes principales, que corresponden tanto a aplicaciones tecnológicas como a teoría psicología (y también a la filosofía de la mente).

La primera es su capacidad para detectar patrones y asociaciones entre patrones cuando se le muestran ejemplos en vez de programarlas de forma explícita.

La segunda es su tolerancia a las pruebas "desordenadas". Mediante la satisfacción de restricciones pueden entender pruebas parcialmente conflictivas. No exigen definiciones rigurosas expresadas como una lista de condiciones necesarias y suficientes, sino que tratan con conjuntos solapados de semejanzas de familia, un rasgo que comparten con los conceptos humanos.

Otra virtud es su capacidad de reconocer patrones incompletos y/o parcialmente dañados. Esto es, tienen memoria de contenido direccionable. Así hacen las personas, por ejemplo, al identificar una melodía con las primeras notas o interpretada con muchos errores.

Y la cuarta, son robustas. Una red de PDP a la que le falten algunos nodos no empieza a producir disparates ni se detiene. Demuestra elegancia en la degradación: el rendimiento empeora gradualmente conforme aumenta el deterioro. Así que no son inestables como los programas simbólicos.

Estos beneficios son consecuencia de la D en PDP. No todas las RNA requieren procesamiento distribuido. En las redes localistas (como WordNet: véase el capítulo II), los conceptos se representan mediante nodos sueltos. En las redes distribuidas, un concepto se almacena (distribuido) por todo el sistema. A veces se combinan la localista y la distribuida, pero no es lo común. Las redes puramente localistas también son poco frecuentes, porque carecen de las virtudes más importantes de los PDP.

Se podría decir que las redes distribuidas son localistas en lo fundamental, ya que cada unidad equivale a un microelemento, por ejemplo, a una mancha minúscula de color en un lugar determinado del campo visual. (No demasiado minúscula y no demasiado determinado: unas cuantas unidades toscamente afinadas son probablemente más eficientes que muchas bien afinadas). Pero las unidades se definen a un nivel mucho más bajo que los conceptos: el PDP requiere computación “sub-simbólica”. Además, cada unidad puede formar parte de muchos patrones generales, así que contribuye a muchos “significados” diferentes.

Hay muchas clases de sistemas de PDP. Todos se componen de tres o más capas de unidades interconectadas, cada una capaz de computar solo una cosa simple. Pero las unidades difieren.

Una unidad de la capa de entrada se dispara cuando a la red se le presenta su microelemento. Una unidad de salida se dispara cuando la activan las unidades conectadas a ella y su actividad se comunica al usuario humano. Las unidades ocultas en la(s) capa(s) intermedia(s) no tienen contacto directo con el mundo exterior. Algunas son deterministas: se disparan, o no, dependiendo solo de

las influencias de sus conexiones. Otras son estocásticas: que se disparen depende en parte de alguna distribución de probabilidad.

Las conexiones también difieren. Algunas son anterógradas y pasan señales de una capa inferior a una superior. Algunas mandan señales retrógradas en sentido opuesto. Algunas son laterales y conectan unidades dentro de una misma capa. Y otras, como veremos, son anterógradas y retrógradas. Igual que las sinapsis cerebrales, las conexiones son o excitadoras o inhibitorias y varían de fuerza o peso. El peso se expresa con números entre +1 y −1. Cuanto mayor sea el peso de una conexión excitadora (o inhibitoria), mayor (o menor) será la probabilidad de que la unidad que recibe la señal se dispare.

El PDP requiere representación distribuida, ya que cada concepto se representa mediante el estado de toda la red. Esto puede parecer desconcertante, incluso paradójico. Desde luego, es muy diferente a cómo se definen las representaciones en la IA simbólica.

A los que solo les interesan las aplicaciones tecnológicas o comerciales no les importa. Si están conformes con que ciertas preguntas obvias (por ejemplo, cómo una sola red puede almacenar varios conceptos o patrones distintos) no sean problemáticas en la práctica, no les preocupa dejar las cosas así.

Los interesados en las implicaciones psicológicas y filosóficas de la IA también hacen esa “pregunta obvia”. La respuesta es que los estados posibles de una red PDP en su conjunto son tan variopintos que solo unos pocos supondrán que esta o aquella distribución de unidades se active simultáneamente. Una unidad activada propagará la activación solo a algunas de las otras unidades. No obstante, esas “otras unidades” varían: cualquier unidad puede participar en muchos patrones de activación diferentes. (En general, las representaciones “dispersas”, con muchas unidades inactivas, son más eficientes). El sistema al final se satura: la investigación teórica sobre las memorias asociativas cuestiona cuántos patrones

pueden almacenar, en principio, las redes de un tamaño determinado.

Pero a los que les interesan los aspectos psicológicos y filosóficos sí les preocupa dejar las cosas así. Les interesa también el concepto de representación en sí mismo y en los debates sobre si la mente / cerebro humanos contienen representaciones internas en realidad.[7] Los adeptos del PDP sostienen, por ejemplo, que este enfoque refuta la hipótesis del sistema de símbolos físicos, que se originó en la IA simbólica y que se extendió rápidamente a la filosofía de la mente (véase el capítulo VI).

EL APRENDIZAJE EN LAS REDES NEURONALES

La mayoría de las RNA pueden aprender. Esto requiere realizar cambios adaptativos en los pesos y a veces también en las conexiones. Por lo general, la anatomía de la red (el número de unidades y los enlaces entre ellas) es fija. En tal caso, el aprendizaje altera solamente los pesos, pero a veces el aprendizaje –o la evolución (véase el capítulo V)– puede añadir conexiones nuevas y reducir conexiones antiguas. Las redes constructivas llevan esto al extremo: empiezan con ninguna unidad oculta y las van añadiendo conforme avanza el aprendizaje.

Las redes de PDP pueden aprender de muchas maneras diferentes y ejemplifican todos los tipos de aprendizaje que hemos distinguido en el capítulo II: aprendizaje supervisado, no supervisado y por refuerzo.

En el aprendizaje supervisado, por ejemplo, llegan a reconocer una clase cuando se le muestran varios ejemplos de ella, ninguno de los cuales tiene que poseer todos los rasgos “típicos”. (Los datos de entrada pueden ser imágenes, descripciones verbales, conjuntos de números…). Cuando se les presenta un ejemplo, algunas unidades de entrada reaccionan a “sus” microelementos y las activaciones se extienden hasta que la red se asienta. El estado resultante de las unidades de salida se compara entonces con el

resultado deseado (determinado por el usuario humano) y se instigan cambios de peso posteriores (quizá mediante retropropagación) para hacer que esos errores sean menos probables. Después de muchos ejemplos que difieran levemente unos de otros, la red habrá desarrollado un patrón de activación que se corresponda con el caso típico o “prototipo”, incluso cuando no se haya topado con ningún caso semejante. (Si en ese momento se le presenta un ejemplo dañado, que estimulará a muchas menos unidades de entrada relevantes, este patrón se completará de manera automática).

La mayoría de los aprendizajes de las RNA se basan en la regla las células que se disparan juntas permanecerán conectadas, que estableció en la década de 1940 el neuropsicólogo Donald Hebb. El aprendizaje hebbiano refuerza las conexiones que se usan con más frecuencia. Cuando dos unidades conectadas entre sí se activan simultáneamente, los pesos se ajustan para que sea más probable que eso vuelva a suceder en el futuro.

Hebb expresó esta regla de dos formas, que no eran ni precisas ni equivalentes. Los investigadores actuales de IA la definen de muchas maneras,[8] basadas quizá en ecuaciones diferenciales sacadas de la física o de la teoría de probabilidad bayesiana. Utilizan el análisis teórico para comparar y mejorar las diferentes versiones. Así, la investigación sobre PDP puede ser endiabladamente matemática. Por eso muchos de los mejores licenciados en física y matemáticas trabajan en instituciones financieras y por eso tan pocos de sus colegas de los centros bursátiles entienden realmente lo que hacen sus sistemas.

Dado que una red de PDP usa alguna regla de aprendizaje hebbiana para adaptar sus pesos, ¿cuándo se detiene? La respuesta no es cuando ha alcanzado la perfección (y se han eliminado todas las inconsistencias), sino cuando ha alcanzado la máxima coherencia.

Una “inconsistencia” tiene lugar, por ejemplo, cuando dos microelementos que no suelen presentarse juntos son señalados simultáneamente por las unidades relevantes. Muchos programas simbólicos de IA pueden resolver problemas de satisfacción de restricciones, acercándose a la solución a base de eliminar las contradicciones entre las pruebas, pero no toleran la inconsistencia como parte de la solución. Los sistemas de PDP son diferentes. Como demuestran las virtudes del PDP enumeradas antes, se pueden desempeñar con éxito aunque persistan las discrepancias. Su “solución” es el estado global de la red cuando se han minimizado (no suprimido) las inconsistencias.

Una forma de conseguir esto pasa por adoptar la idea del equilibrio de la termodinámica. Los niveles de energía en física se expresan numéricamente, igual que los pesos en PDP. Si la regla de aprendizaje es similar a las leyes físicas (y si las unidades ocultas son estocásticas), las mismas ecuaciones estadísticas de Boltzmann pueden describir los cambios en ambos casos.

El PDP puede incluso adoptar el método que se utiliza para enfriar metales rápida pero uniformemente. El recocimiento de metales empieza a una temperatura alta y se va enfriando gradualmente. Los que investigan sobre PDP a veces usan un cocimiento simulado en el que los cambios de peso en los primeros ciclos para equilibrar son mucho mayores que los de los últimos ciclos. Esto permite que la red escape de situaciones (“mínima local”) en las que la consistencia global se ha alcanzado en relación a lo que sucedió antes, pero podría conseguirse una consistencia aún mayor (y un equilibrio más estable) si el sistema fuese perturbado. Podemos compararlo con agitar una bolsa de canicas para sacar una que se haya quedado en un pliegue interior: habría que empezar por sacudir con fuerza, pero terminar agitando con suavidad.

Otra forma más rápida y más utilizada de alcanzar la consistencia máxima consiste en emplear la retropropagación, pero sea cual sea la regla de aprendizaje que se utilice, el estado de la

red completa (y especialmente de las unidades de salida) en equilibrio se toma como representación del concepto en cuestión.

RETROPROPAGACIÓN Y CEREBRO… Y APRENDIZAJE PROFUNDO

Los entusiastas de la retropropagación y del aprendizaje profundo PDP sostienen que sus redes son biológicamente más realistas que la IA simbólica. Es cierto que el PDP está inspirado en el cerebro y que algunos neurocientíficos lo usan para replicar el funcionamiento neuronal. Sin embargo, las RNA difieren significativamente de lo que llevamos dentro de la cabeza.

Una diferencia entre (la mayoría de) las RNA y el cerebro es la retropropagación. Es una regla de aprendizaje (o, más bien, una clase general de reglas de aprendizaje) que se usa con frecuencia en el PDP. Anticipada por Paul Werbos en 1974, Geoffrey Hinton la definió de manera más útil a principios de la década de 1980.[9] Con ella se resuelve el problema de asignación del mérito.

Este problema surge en todos los tipos de IA, especialmente cuando el sistema está cambiando continuamente. En un sistema complejo de IA dado que logre su fin, ¿qué partes son responsables del logro? En la IA evolutiva, el mérito se suele asignar mediante el algoritmo de “bucket-brigade” (véase el capítulo V). En los sistemas PDP con unidades deterministas (no estocásticas), el mérito se asigna normalmente mediante retropropagación.

El algoritmo de retropropagación determina el origen de la responsabilidad desde la capa de salida hasta las capas ocultas, identificando las unidades individuales que necesitan ser ajustadas. (Los pesos se actualizan para minimizar los errores de predicción). El algoritmo necesita conocer el estado preciso de la capa de salida cuando la red esté dando la respuesta correcta. (Por tanto, la retropropagación es un aprendizaje supervisado). Se realizan comparaciones unidad por unidad entre este resultado ejemplar y los obtenidos realmente por la red. Cualquier diferencia con la

actividad de una unidad de salida en los dos casos cuenta como error.

El algoritmo asume que el error en una unidad de salida se debe a error(es) en las unidades conectadas con ella. Al trabajar hacia atrás por todo el sistema, le atribuye una cantidad específica del error a cada unidad en la primera capa oculta, según el peso de la conexión entre ella y la unidad de salida. La culpa se comparte entre todas las unidades ocultas conectadas a la unidad de salida equivocada. (Si una unidad oculta está conectada con varias unidades de salida, sus mini-culpas se suman). Entonces se realizan cambios de peso proporcionales en las conexiones entre la capa oculta y la capa precedente.

Esa capa puede ser otro (y otro…) estrato de unidades ocultas, pero en última instancia será la capa de entrada y los cambios de peso se detendrán. Este proceso se itera hasta que las discrepancias en la capa de salida se minimizan.

Durante muchos años, la retropropagación se utilizaba solo en redes con una capa oculta. Las redes multicapa eran escasas: es difícil analizarlas y hasta experimentar con ellas. En los últimos tiempos, sin embargo, han creado un interés tremendo (y alguna propaganda irresponsable) por el advenimiento del aprendizaje profundo,[10] en el que un sistema aprende la estructura penetrando hasta lo profundo de un campo, en oposición a los meros patrones superficiales. Dicho de otro modo, descubre una representación del conocimiento de varios niveles, no de un solo nivel.

El aprendizaje profundo es interesante porque promete que las RNA, por fin, podrán manejar la jerarquía. Desde principios de la década de 1980, los conexionistas como Hinton y Jeff Elman se esforzaron por representar la jerarquía combinando la representación local / distribuida o definiendo las redes recurrentes. (Las redes recurrentes, en efecto, funcionan como una secuencia de pasos discretos. Las versiones recientes que usa el aprendizaje profundo pueden predecir a veces la siguiente palabra de una frase

o incluso el “pensamiento” siguiente de un párrafo).[11] Pero tuvieron un éxito limitado (y las RNA no son todavía adecuadas para representar con precisión jerarquías definidas ni razonamiento deductivo).

El aprendizaje profundo, también, empezó en la década de 1980 (con Jurgen Schmidhuber). Pero el campo estalló hace mucho menos, cuando Hinton fue capaz de brindar un método eficiente que permitía que las redes multicapa descubriesen relaciones en muchos niveles.[12] Sus sistemas de aprendizaje profundo están compuestos por máquinas de Boltzmann “restringidas” (sin conexiones laterales) en media docena de capas. Primero, las capas llevan a cabo un aprendizaje no supervisado. Se entrenan una por una mediante una simulación de cocimiento. La salida de una capa se usa como entrada de la siguiente. Cuando la última capa se estabiliza, todo el sistema habrá sido ajustado mediante retropropagación llegando hasta abajo a través de todos los niveles para asignar méritos como corresponde.

Este enfoque del aprendizaje es interesante para los neurocientíficos cognitivos y para los tecnólogos de la IA, porque establece “modelos generativos” que aprenden a predecir las causas (probables) de las entradas en la red, proporcionando así un modelo de lo que Helmholtz llamó en 1867 “percepción como inferencia inconsciente”. Es decir, que la percepción no es cuestión de recibir estímulos de manera pasiva desde los órganos de los sentidos: requiere de interpretación activa y hasta de predicción anticipatoria de esos estímulos. En suma, el sistema ojo / cerebro no es una cámara.

Hinton se incorporó a Google en 2013, así que la retropropagación va a estar muy ocupada. Google ya utiliza el aprendizaje profundo en muchas aplicaciones, entre ellas el reconocimiento de voz y el procesamiento de imágenes. Además, en 2014 compró DeepMind, cuyo algoritmo DQN dominó los juegos clásicos de Atari combinando el aprendizaje profundo con el aprendizaje por refuerzo (véase el capítulo II). IBM también le está

dando prioridad al aprendizaje profundo: WATSON lo utiliza y muchas aplicaciones especializadas están en proceso de adoptarlo (véase el capítulo III).

Sin embargo, aunque el aprendizaje profundo sea indiscutiblemente útil, no significa que se entienda perfectamente. Se están estudiando experimentalmente muchas reglas de aprendizaje multicapa diferentes, pero el análisis teórico es confuso. Entre las innumerables preguntas sin responder está la de si hay profundidad suficiente para alcanzar un resultado casi humano. La unidad que reconoció caras de gatos mencionada en el capítulo II resultó de un sistema de nueve capas: pero ¿cuántas capas se añaden mediante cálculos en el córtex cerebral? Dado que las RNA se inspiran en el cerebro (como nos destacan constantemente en la promoción del aprendizaje profundo), esa pregunta es natural, pero no tan pertinente como pueda parecer.

La retropropagación es un triunfo de la informática, pero es no es nada biológica. Ninguna "célula abuela" reconocedora de caras de gatos (véase el capítulo II) podría originarse en el cerebro a partir de procesos como los del aprendizaje profundo. Las sinapsis reales están exclusivamente alimentadas hacia adelante, no transmiten en ambas direcciones. El cerebro contiene conexiones retroalimentadas en varias direcciones, pero todas son de sentido único. Esta es una de las muchas diferencias entre las redes neuronales reales y artificiales. (Otra es que las redes cerebrales no se organizan como jerarquías estrictas, aunque el sistema visual se suela describir de esa forma).

El hecho de que el cerebro contenga conexiones tanto hacia adelante como hacia atrás es crucial para los modelos de codificación predictiva del control sensoriomotor, que están despertando gran interés en neurociencia. (También están basados en gran parte en el trabajo de Hinton). Los niveles neuronales más altos envían mensajes hacia abajo prediciendo las señales de entrada de los sensores y los únicos que no se envían hacia arriba son los mensajes de "error" imprevistos. Este tipo de ciclos, al

repetirse, van ajustando las redes predictivas para que aprendan gradualmente qué esperar. Los investigadores hablan del “cerebro bayesiano” porque las predicciones se pueden interpretar en términos de estadística bayesiana, que son realmente los términos en los que se basan los modelos informáticos; véase el capítulo II.

Comparadas con el cerebro, las RNA son demasiado claras, demasiado simples, demasiado escasas y demasiado pobres. Demasiado claras, porque las redes de construcción humana priorizan la elegancia matemática y la potencia, mientras que el cerebro evolucionado de manera biológica no. Demasiado simples, porque una sola neurona (de las que hay unas treinta clases diferentes) es tan compleja computacionalmente como un sistema PDP completo o incluso como un pequeño ordenador. Demasiado escasas, porque incluso las RNA con millones de unidades son minúsculas comparadas con el cerebro humano (véase el capítulo VII). Y demasiado pobres porque los investigadores de RNA por lo general ignoran además de los factores temporales como frecuencias neuronales en su punto máximo y sincronías, la biofísica de las espinas dendríticas, los neuromoduladores, las corrientes sinápticas y el paso de iones.

Pero estas limitaciones, todas ellas, se está reduciendo. El aumento de la potencia de los ordenadores está permitiendo que las RNA contengan muchas más unidades individuales. Se están creando modelos de neuronas mucho más detallados que ya se encargan de las funciones computacionales de todos los factores neurológicos que se acaban de mencionar. La “pobreza” está disminuyendo en la realidad, igual que en la simulación (algunas investigaciones “neuromórficas” combinan neuronas vivas con chips de silicio). Y por mucho que el algoritmo DQN simule procesos en el córtex visual y en el hipocampo (véase el capítulo II), las RNA futuras sin duda adoptarán otras funciones de la neurociencia.

Sin embargo, sigue siendo cierto que las RNA no son como el cerebro en infinitos aspectos, algunos de los cuales ni siquiera conocemos todavía.

EL ESCÁNDALO DE LA RED

La emoción por la llegada del PDP se debió en gran parte a que se había declarado la muerte de las RNA (alias conexionismo) veinte años antes. Como se señaló en el capítulo I, ese juicio había llegado en una crítica salvaje de 1960 de Marvin Minsky y Seymour Papert, ambos con reputaciones estelares dentro de la comunidad de la IA. Cuando llegó la década de 1980, las RNA no parecían en punto muerto, sino muertas. De hecho, se había marginado en general a los cibernéticos (véase el capítulo I). Casi todos los fondos de investigación habían ido a parar a la IA simbólica.

Un poco antes, las RNA parecían tremendamente prometedoras. Los perceptrones autoorganizativos de Rosenblatt (muchas veces observados por periodistas fascinados) eran capaces de aprender a reconocer patrones incluso si empezaban desde un estado aleatorio. Rosenblatt había hecho aseveraciones tremendamente ambiciosas que abarcaban toda la psicología humana gracias al potencial de su planteamiento. Había señalado ciertas limitaciones, por supuesto, pero su intrigante "teorema de la convergencia" había garantizado que unos simples perceptrones podrían aprender cualquier cosa siempre que fuera posible programarlos para que las aprendieran. Era mucho decir.

Pero Minsky y Papert, a finales de la década de 1960, presentaron sus propias pruebas.[13] Demostraron matemáticamente que esos simples perceptrones no pueden hacer ciertas cosas que de manera intuitiva se esperaría que hicieran (y que la inteligencia artificial simbólica podría hacer sin dificultad). Sus pruebas (como el teorema de convergencia de Rosenblatt) se aplicaban únicamente a las redes de una sola capa, pero su "juicio intuitivo" era que a los sistemas multicapa los sobrepasaría la explosión combinatoria. Dicho de otro modo, los perceptrones no crecerían.

La mayoría de los científicos que trabajaban en IA estaban convencidos de que el conexionismo no tendría éxito nunca. Unos

pocos siguieron con la investigación sobre RNA a pesar de todo. De hecho, algunos progresos muy significativos se hicieron al analizar la memoria asociativa (Christopher Longuet-Higgins y David Willshaw y luego James Anderson, Teuvo Kohonen y John Hopfield).[14] Pero esos trabajos quedaron en segundo plano. Los grupos en cuestión no se identificaban como investigadores de “IA” y por lo general eran ignorados por los que sí.

La llegada del PDP acabó con ese escepticismo. Además de algunos modelos en marcha impresionantes (como el aprendiz del pretérito), había dos nuevos teoremas convergentes: uno que garantizaba que un sistema con PDP basado en las ecuaciones termodinámicas de Boltzmann alcanzaría el equilibrio (aunque quizá después de mucho tiempo) y otro que demostraba que una red de tres capas podía resolver en principio cualquier problema que se le presentara. (Advertencia sanitaria: igual que en el caso de la IA simbólica, representar un problema de manera que se pueda introducir en el ordenador suele ser la parte más difícil de la tarea). Como es natural, sobrevino el entusiasmo. El consenso sobre la IA dominante se hizo añicos.

La IA simbólica había asumido que el pensamiento intuitivo espontáneo es como la inferencia consciente, pero sin la consciencia. Ahora, los investigadores del PDP decían que eran dos clases de pensamiento fundamentalmente distintas. Todos los líderes del movimiento de PDP (David Rumelhart, Jay McClelland, Donald Norman y Hinton) señalaron que ambos tipos son clave para la psicología humana. Pero la propaganda sobre el PDP (y la reacción del público en general) implicaba que la IA simbólica, considerada como el estudio de la mente, era una pérdida de tiempo. Había llegado la hora de la revancha.

El principal proveedor de fondos de la IA, el departamento de Defensa de Estados Unidos, también dio un giro de 180º. Después de una reunión urgente en 1988, admitió que su negligencia anterior de las RNA no había sido “merecida”. Entonces, a la investigación sobre PDP le llovió el dinero.

En cuanto a Minsky y Papert, fueron contumaces.[15] Concedían que “la riqueza del futuro del aprendizaje automático basado en redes excede la imaginación”, pero insistían en que no puede surgir una inteligencia de alto nivel del puro azar ni de un sistema completamente no secuencial. En consecuencia, el cerebro debe actuar a veces como un procesador en serie y la IA de nivel humano tendrá que emplear sistemas híbridos. Arguyeron que su crítica no había sido el único factor que había conducido a las RNA a sus años de sequía: en primer lugar, la potencia de los ordenadores era insuficiente. Y negaron que hubiesen estado intentando desviar el dinero para la investigación a la IA simbólica. Según dijeron, “No pensamos que nuestro trabajo estuviera matando a Blancanieves, sino que era una forma de entenderla”.

Estos eran argumentos científicos respetables, pero su crítica inicial rezumaba veneno. (El borrador era todavía más venenoso: algunos colegas los persuadieron de que le rebajaran el tono para darle más importancia a los puntos científicos). No tiene nada de raro que desencadenara emociones. Los perseverantes partidarios de las RNA se tomaron con profundo resentimiento su recién hallada invisibilidad cultural. El furor causado por el PDP fue aún mayor. La “muerte” y el renacimiento de las RNA pasó por episodios de celos, desprecio, auto ensalzamiento y regodeo burlón: “¡Os lo dijimos!”.

Este episodio fue un claro ejemplo de escándalo científico y no el único que surgió dentro de la IA.[16] Los desacuerdos teóricos se embrollaron con sentimientos personales y rivalidades y la imparcialidad escaseaba. Amargos insultos llenaron el aire y las imprentas también. La IA no es una aventura desapasionada.

LAS CONEXIONES NO LO SON TODO

La mayoría de los estudios sobre las RNA sugieren que lo único que importa de la red neuronal es su anatomía. ¿Qué unidades están conectadas a qué otras y cómo de fuertes son los pesos?

Cierto, estas cuestiones son cruciales. Sin embargo, la neurociencia ha demostrado recientemente que los circuitos biológicos a veces pueden alterar su función computacional (no solo hacer que sea más o menos probable) gracias algunos compuestos químicos que se difunden por el cerebro.

El monóxido de nitrógeno (NO), por ejemplo, se difunde en todas direcciones y sus efectos (que dependen de su concentración en los puntos relevantes) perduran hasta que decae su concentración. (Las enzimas pueden cambiar la tasa de desintegración). Así, el NO trabaja en todas las células que se encuentran en un volumen determinado del córtex, ya estén conectadas sinápticamente o no. La dinámica funcional de los sistemas neuronales en cuestión es muy diferente de las RNA "puras", ya que la comunicación por volumen reemplaza a la comunicación punto por punto. Se han descubierto efectos análogos con el monóxido de carbono y el sulfuro de hidrógeno y con moléculas complejas como la serotonina y la dopamina.

"¡Hasta aquí pueden llegar las RNA! –podría decir un escéptico de la IA–. ¡Dentro de los ordenadores no hay química!". Este comentario es absurdo: es como decir que los ordenadores no pueden imitar el clima porque en su interior no puede llover. "Por lo tanto –podrían añadir–, la IA no puede crear modelos de estados de ánimo o emociones, ya que estos dependen de las hormonas y de los neuromoduladores". Esta misma objeción fue expresada por el psicólogo Ulric Neisser a principios de la década de 1960[17] y unos años después por el filósofo John Haugeland en su influyente crítica al "cognitivismo".[18] La IA puede crear modelos del razonamiento, dijeron, pero nunca del afecto.

Sin embargo, estos descubrimientos neurocientíficos han inspirado a algunos investigadores de IA a diseñar RNA de un tipo radicalmente nuevo, en las que la conexión no lo es todo.[19] En GasNets, algunos nodos diseminados por la red pueden soltar "gases" simulados que se difunden y modulan las propiedades intrínsecas de otros nodos y conexiones de diferentes formas según

su concentración. El volumen de la difusión importa, igual que la forma de la fuente de origen (modelada como una esfera hueca, no como una fuente puntual). Así, un módulo dado se comportará de forma diferente en momentos diferentes. En ciertas condiciones gaseosas, un nodo afectará a otro a pesar de que no exista una conexión directa entre ellos. Lo crucial es la interacción entre el gas y las conectividades eléctricas dentro del sistema. Y, como el gas se emite solo en ciertas ocasiones y se difunde y desintegra a tasas variables, esta interacción es dinámicamente compleja.

La tecnología GasNet se utilizó, por ejemplo, para desarrollar "cerebros" destinados a robots autónomos. Los investigadores descubrieron que un comportamiento específico podía involucrar a dos sub-redes no conectadas, que trabajaban juntas debido a los efectos modulatorios. Descubrieron también que un "detector de orientación" capaz de usar un triángulo de cartón como referencia para desplazarse podía adoptar la forma de sub-redes parcialmente no conectadas. Previamente habían elaborado una red completamente conectada para hacerlo (véase el capítulo V), pero la versión neuromoduladora fue desarrollada más rápido y era más eficiente.

Así que algunos investigadores de RNA han pasado de tener en cuenta solo la anatomía (conexiones) a tener en cuenta también la neuroquímica. Ahora se pueden simular diferentes normas de aprendizaje y sus interacciones temporales teniendo la neuromodulación en mente.

La neuromodulación es un fenómeno analógico, no digital. Las concentraciones de moléculas disueltas, en constante variación, son importantes. Los investigadores de IA (utilizando chips VLSI especiales) diseñan cada vez más redes que combinan funciones analógicas y digitales. Los elementos analógicos se replican con la anatomía y la fisiología de las neuronas biológicas, incluyendo el paso de iones a través de la membrana celular. Esta computación "neuromórfica" se está utilizando, por ejemplo, para simular aspectos de la percepción y del control motor. Algunos científicos de

IA planean utilizar la computación neuromórfica dentro del modelo del “cerebro completo” (véase el capítulo VII).

Otros van más lejos aún: en vez de replicar las RNA únicamente in silico, construyen (o desarrollan: véase el capítulo V) redes compuestas de electrodos en miniatura y neuronas reales. Por ejemplo, cuando los electrodos X e Y se estimulan artificialmente, la actividad resultante en la red de “wetware” provoca que se active algún otro electrodo, Z, y así se aplique una compuerta AND. Este tipo de computación (concebida por Donald Mackay en la década de 1940)[20] está en pañales, pero es potencialmente fascinante.

SISTEMAS HÍBRIDOS

Las redes analógicas / digitales y de hardware / wetware que se acaban de mencionar podrían describirse de manera comprensible como sistemas “híbridos”, pero este término se usa normalmente para referirse a programas de IA que incluyen tanto el procesamiento de la información simbólico como el conexionista.

Minsky, en su manifiesto,[21] dijo que probablemente ambos eran necesarios y al principio algunos programas simbólicos combinaban procesamiento secuencial y paralelo, pero no había muchos de esos intentos. Como se ha visto, Minsky siguió recomendando híbridos simbólicos / RNA después de la llegada del PDP. Sin embargo, ese tipo de sistemas no siguieron de inmediato (aunque Hinton construyó redes combinando conexionismo localista y distribuido para representar jerarquías todo / parte como árboles genealógicos).

De hecho, la integración de procesamiento simbólico y procesamiento mediante redes neuronales sigue siendo poco común. Las dos metodologías, la lógica y la probabilística, son tan diferentes que la mayoría de los investigadores se especializan solo en una.

Sin embargo, se han desarrollado algunos sistemas genuinamente híbridos, en los que el control pasa entre módulos

simbólicos y módulos de procesamiento distribuido en paralelo según proceda. Así, el modelo se basa en las virtudes complementarias de ambos enfoques.

Entre los ejemplos, están los algoritmos para jugar desarrollados por DeepMind (véase el capítulo II) que combinan aprendizaje profundo con inteligencia artificial simbólica para aprender a jugar a diversos juegos de ordenador. Utilizan el aprendizaje por refuerzo: no se les proporcionan reglas manufacturadas, solo los píxeles de entrada y las puntuaciones numéricas de cada paso. Se consideran muchas reglas / planes de manera simultánea y los más prometedores deciden cuál será la acción siguiente. (Las versiones futuras se concentrarán en juegos en 3D como Minecraft y en aplicaciones como los vehículos sin conductor).

Otros ejemplos son los sistemas cognitivos completos ACT-R* y CLARION (véase el capítulo II) y LIDA (véase el capítulo VI), basados fundamentalmente en la psicología cognitiva y que han sido desarrollados con propósitos científicos, no tecnológicos.

Algunos modelos híbridos también tienen en cuenta aspectos específicos de la neurología.[22] Por ejemplo, el neurólogo clínico Timothy Shallice, junto al pionero del PDP Norman, publicó una teoría híbrida sobre acciones conocidas ("sobreaprendizaje") en 1980, que se implementó más tarde.[23] Esta teoría explica algunos errores comunes. Por ejemplo, los pacientes con apoplejía se suelen olvidar de que una carta debería meterse en el sobre antes de pegar la solapa; o pueden meterse en la cama cuando suben a cambiarse de ropa, o coger un cazo en vez de la tetera. Errores similares –de orden, captura y sustitución de objetos– nos ocurren a todos de vez en cuando.

Pero ¿por qué? ¿Y por qué los pacientes con daño cerebral son especialmente propensos a ellos? La teoría computacional de Shallice afirma que una acción conocida es generada por dos tipos de control que pueden descomponerse o hacerse con el mando en momentos específicos. Uno, la "selección y aplicación de rutinas",

es automática. Requiere competición (inconsciente) entre diferentes esquemas de acciones organizadas jerárquicamente. El control es para aquel cuya activación haya superado cierto umbral. El otro mecanismo de control (“ejecutivo”) es consciente. Requiere supervisión deliberativa y modulación del primer mecanismo, incluyendo planificación y subsanación de errores. Para Shallice, la selección y aplicación de rutinas se replica mediante PDP y la función ejecutiva mediante IA simbólica.

Se puede aumentar el nivel de activación de un esquema de acción mediante entradas perceptuales. Por ejemplo, un atisbo impensado (reconocimiento de patrones) de la cama al llegar al dormitorio puede desencadenar el esquema de acción de meterse en ella, aunque la intención original (el plan) fuese cambiarse de ropa.

La teoría de la acción de Shallice empezó utilizando ideas de la IA (sobre todo, modelos de planificación) que se hacían eco de su propia experiencia clínica, respaldada posteriormente con pruebas conseguidas mediante escáneres cerebrales. Y la neurociencia reciente ha descubierto otros factores, como los neurotransmisores, implicados en la acción humana. Ahora se representan en los modelos computacionales actuales basados en la teoría.[24]

Las interacciones entre la selección y aplicación de rutinas y la función ejecutiva son relevantes también para la robótica. Un agente que siga un plan debería poder detenerlo o variarlo según lo que observe en el medio. Esa estrategia caracteriza a los robots que combinan procesamiento situado y deliberativo (véase el capítulo V).

Cualquiera interesado en la IAF debería tener en cuenta que los pocos científicos de IA que han considerado seriamente la arquitectura computacional de la mente como un todo aceptan el hibridismo sin reservas. Entre ellos figuran Allen Newell y Anderson (de cuyos soar y ACT* se habló en el capítulo II), Stan Franklin (cuyo modelo de consciencia LIDA se describe en el capítulo VI), Minsky (con su teoría de la “sociedad” de la mente)[25] y Aaron

Sloman (cuya simulación de la ansiedad se describe en el capítulo III).

En suma, las máquinas virtuales implementadas en la mente humana son tanto secuenciales como paralelas. La inteligencia humana requiere que ambas cooperen de manera sutil, y la IAF a nivel humano, si alguna vez se consigue, también lo hará.

V

LOS ROBOTS Y LA VIDA ARTIFICIAL

La vida artificial (A-Life) imita a los sistemas biológicos. Como la IA en general, tiene objetivos tecnológicos y científicos.[1] La vida artificial es fundamental para la IA, ya que toda la inteligencia de la que tenemos noticia se da en organismos vivos. De hecho, hay quien cree que la mente surge solo si hay vida (véase el capítulo VI). Esa cuestión no les preocupa a los tecnólogos testarudos; lo que hacen es recurrir a la biología para desarrollar aplicaciones prácticas de muchas clases, como los robots, la programación evolutiva y los dispositivos autoorganizados. Los robots son la quintaesencia de la IA: tienen gran repercusión, son tremendamente ingeniosos y representan un gran negocio. La IA evolutiva, aunque de uso generalizado, es poco conocida, y las máquinas autoorganizadas todavía menos (excepto el aprendizaje no supervisado: véase el capítulo IV). No obstante, en el proceso para comprender la autoorganización, la IA le ha sido tan útil a la biología como la biología a la IA.

ROBOTS SITUADOS E INSECTOS INTERESANTES

Los robots se empezaron a construir hace siglos, por parte de Leonardo da Vinci, entre otros; los modelos con IA aparecieron en la década de 1950. Las “tortugas” que William Grey Walter presentó en la posguerra asombraron a los observadores porque evitaban obstáculos y encontraban la luz, y un objetivo principal del entonces recién fundado laboratorio de IA del MIT (Instituto Tecnológico de Massachusetts) era construir “el robot MIT”, que incorporase visión artificial, planificación, lenguaje y control motor.

Se ha avanzado muchísimo desde entonces. Ahora, algunos robots pueden subir pendientes, escaleras o muros; otros pueden correr con rapidez o saltar alto; otros pueden cargar (y arrojar) cargas pesadas, desmontarse a sí mismos y reensamblar las partes, a veces adoptando formas nuevas (de gusano capaz de recorrer una tubería estrecha, o de pelota o criatura con muchas patas adecuadas para terrenos llanos o accidentados respectivamente). Lo que impulsó este adelanto fue el cambio de la psicología a la biología.

Los robots de la IA clásica emulaban la voluntad humana; inspirándose en teorías de replicación cerebral, empleaban las representaciones internas del mundo y de las acciones del propio agente, pero no eran muy impresionantes. Como se basaban en planificación abstracta, estaban sujetos al problema del marco (véase el capítulo II). No eran capaces de reaccionar de inmediato, porque hasta con leves cambios en el medio requerían planificación anticipatoria para reanudarse; tampoco se podían adaptar a circunstancias nuevas (no modeladas). Les era difícil moverse con estabilidad incluso en terreno nivelado y despejado (de ahí el mote del robot SRI, SHAKEY, tembloroso) y los robots caídos no eran capaces de recuperarse. Eran inútiles en la mayoría de los edificios, no hablemos ya en Marte.

Los robots actuales son muy diferentes. Los insectos han sustituido a los seres humanos como el centro de atención. Probablemente los insectos no sean lo suficientemente inteligentes como para imitar el mundo o para planificar, pero se las arreglan. Su comportamiento (comportamiento, no acción) es apropiado, adaptativo, pero en lo fundamental es un reflejo más que deliberado. Los insectos reaccionan de manera irreflexiva a la situación del momento, no a alguna posibilidad imaginada o estado objetivo, de ahí las etiquetas de robótica “situada” o “basada en el comportamiento”. (El comportamiento situado no se circunscribe a los insectos: los psicólogos sociales han identificado en los seres humanos muchos comportamientos vinculados a la situación).

A la hora de intentar darles a las máquinas con IA reflejos comparables a los de los insectos, los expertos en robótica prefieren la ingeniería a la programación. Cuando es posible, los reflejos sensoriomotores no se incorporan en la anatomía del robot como código informático, sino físicamente.

Es discutible hasta dónde debería concordar la anatomía robótica con la anatomía de los organismos vivos. Para propósitos tecnológicos, los trucos de ingeniería ingeniosos son aceptables. Los robots actuales incorporan muchos artilugios poco realistas, pero ¿no será que los mecanismos biológicos son especialmente eficientes? Apropiados son, sin duda, por eso los expertos en robótica también estudian animales reales: lo que pueden hacer (incluyendo sus distintas estrategias de orientación), las señales sensoriales y los movimientos específicos que usan para ello y los mecanismos neurológicos responsables. Los biólogos, por su parte, emplean modelos de IA para investigar estos mecanismos, en un campo de investigación llamado neuroetología computacional.

Un ejemplo son las cucarachas robóticas de Randall Beer,[2] que tienen seis patas multisegmentadas, lo que ofrece ventajas y desventajas. La locomoción de los hexápodos es más estable que la de los bípedos (y por lo general más útil que las ruedas). No obstante, coordinar seis extremidades parece más difícil que coordinar dos. Además de decidir qué pata habría que mover a continuación, la criatura tiene que averiguar el emplazamiento, la fuerza y la ocasión correctos. ¿Y cómo deberían interactuar las patas entre sí? Tendrían que ser muy independientes, porque podría haber un guijarro cerca de una de ellas, pero, si esta se levanta de más, las otras tendrían que compensarse para conservar el equilibrio.

Los robots de Beer reflejan la neuroanatomía y los controles sensoriomotores de las cucarachas de verdad. Pueden subir escaleras, andar por terrenos accidentados, trepar por encima de los obstáculos (en vez de limitarse a evitarlos) y recuperarse de una caída.

Barbara Webb no se fija en las cucarachas sino en los grillos.[3] No le interesa la locomoción (así que sus robots pueden usar ruedas): quiere que sus dispositivos identifiquen y localicen un patrón de sonido particular y sean capaces de dirigirse hacia él. Queda claro que ese comportamiento (“fonotaxis”) podría tener muchas aplicaciones prácticas.

Las hembras de los grillos se comportan así al oír el canto de un macho conespecífico. No obstante, solo pueden reconocer un canto que tenga una velocidad y una frecuencia en particular, que varían según las distintas especies de grillos. Pero la hembra no elige entre diferentes cantos, porque no posee detectores de rasgos que codifiquen un rango de sonidos; utiliza un mecanismo sensible a una sola frecuencia que no es neuronal, como los detectores auditivos del cerebro humano, sino un tubo de longitud fija situado en el tórax, conectado a los oídos de las patas delanteras y los espiráculos. La longitud del tubo está en proporción exacta con la longitud de onda del canto del macho. Las leyes físicas garantizan que las cancelaciones de fase (entre el aire y el tubo y el aire exterior) tengan lugar solo cuando haya cantos con la frecuencia adecuada y la diferencia de intensidad dependa completamente de la dirección de la fuente del sonido. El insecto hembra está programado neuronalmente para moverse en esa dirección: él canta, ella acude. Comportamiento situado, en efecto.

Webb eligió la fonotaxis de los grillos porque los neuroetólogos la habían estudiado de cerca, pero quedaban muchas preguntas sin responder: si (y cómo) la dirección y el sonido del canto se procesan por separado; si la identificación y la localización son independientes; cómo se desencadena el desplazamiento de la hembra; y cómo se controla la orientación en zigzag. Webb diseñó el mecanismo más simple posible (con solo cuatro neuronas) capaz de generar el mismo comportamiento. Luego su modelo usaría más neuronas (basadas en real life data), incluiría rasgos neuronales adicionales (como latencia, tasa de disparo y potencial de membrana) y aunaría el oído con la visión. Su trabajo ha aclarado

muchas cuestiones neurocientíficas, respondiendo otras y planteando más, así que ha sido útil tanto para la biología como para la robótica.

(Aunque los robots son cosas físicas, gran parte de la investigación robótica se hace mediante simulación. A los robots de Beer, por ejemplo, a veces se les hace evolucionar en software antes de construirlos. De la misma forma, se diseñan como programas antes de probarlos en el mundo real).

A pesar de que la corriente principal de la robótica se haya desviado hacia los insectos, la investigación sobre robots androides prosigue. Algunos son meros juguetes, otros son robots “sociales” o “de compañía” de uso doméstico diseñados para personas ancianas y/o incapacitadas (véase el capítulo III). Su finalidad no es tanto la de esclavos recaderos como de asistentes personales autónomos. Algunos tienen “buena” apariencia, con largas pestañas y voces seductoras. Pueden mirar a los ojos a los usuarios y reconocer caras y voces individuales. También pueden (hasta cierto punto) mantener conversaciones improvisadas, interpretar el estado emocional del usuario y generar reacciones “emocionales” (expresiones faciales y/o patrones de habla parecidos a los humanos).

Aunque algunos robots son grandes (para manejar cargas pesadas y/o atravesar terrenos accidentados), no es lo más habitual, y existen algunos (lo que se usan dentro de los vasos sanguíneos, por ejemplo) realmente pequeños. Por lo general trabajan en grandes grupos. Cuando hay múltiples robots involucrados en una tarea, surgen interrogantes sobre cómo (si es que acaso) se comunican y cómo esa comunicación le permite al grupo hacer cosas que no podrían hacer sus miembros individualmente.

Para encontrar respuestas, los expertos en robótica suelen volver la vista a los insectos sociales, como las hormigas y las abejas. Este tipo de especies ejemplifican la “cognición distribuida”

(véase el capítulo II), en la que el conocimiento (y las acciones pertinentes) se disemina por todo el grupo en vez de ser asequible para cualquier animal como individuo.

Si los robots son muy sencillos, sus desarrolladores pueden hablar de "inteligencia de enjambre" y analizar sistemas robóticos cooperativos, como los autómatas celulares (AC). Un autómata celular es un sistema de unidades individuales que adoptan un número finito de estados siguiendo reglas sencillas que dependen del estado de sus vecinas. El patrón general de comportamiento de un autómata celular puede ser sorprendentemente complejo. La analogía básica son las células vivas que cooperan en organismos multicelulares. Las muchas versiones de IA incluyen los algoritmos de movimiento en bandada que se usan para los murciélagos o los dinosaurios de las animaciones de Hollywood.

Los conceptos de cognición distribuida e inteligencia de enjambre también se aplican a los seres vivos. Esta última es la que se usa cuando los "conocimientos" en cuestión se refieren a algo que ningún participante individual puede tener (por ejemplo, el comportamiento global de grandes multitudes). La primera se utiliza más cuando los individuos participantes podrían tener todos los conocimientos relevantes, pero no los tienen. Por ejemplo, un antropólogo ha demostrado cuántos conocimientos de navegación comparten los miembros de la tripulación de un barco y se materializan también en objetos físicos, como mapas y (la localización de las) cartas de navegación.[4]

Hablar de conocimiento materializado en objetos físicos puede parecer extraño o, en el mejor de los casos, metafórico, pero hoy son muchos los que aseguran que la mente humana es literalmente una encarnación, no solo en las acciones físicas de la gente sino también en los artefactos culturales con los que se relacionan en el mundo exterior. Esta teoría de "mente externa / encarnada" se basa en parte en el trabajo realizado por quien lideró el paso de los humanos a los insectos en la robótica: Rodney Brooks, del MIT.

Brooks es ahora un destacado desarrollador de robots para el ejército de Estados Unidos. En la década de 1980, era un experto en robótica en ciernes frustrado por la impracticabilidad de los planificadores de modelos científicos de IA simbólica. Se pasó a los robots situados por razones puramente tecnológicas, pero pronto adoptó un método para realizar una teoría sobre comportamiento adaptivo en general.[5] Esto iba más allá de los insectos: ni siquiera los actos humanos, afirmó, requieren representaciones internas. (O, como dio a entender a veces, no suelen requerir representaciones).

Su crítica de la IA simbólica entusiasmó a psicólogos y filósofos, algunos de los cuales se mostraron muy partidarios. Los psicólogos ya habían señalado que gran parte del comportamiento humano está ligado a la situación, como los roles que se representan en distintos círculos sociales, por ejemplo. Y los psicólogos cognitivos habían destacado la visión animada, en la que es clave el movimiento corporal del propio agente. Hoy, las teorías de cognición corpórea son muy influyentes fuera de la IA (véase el capítulo VI).

Pero otros, como David Kirsh,[6] se oponían (y se oponen) con vehemencia, arguyendo que las representaciones composicionales son necesarias para aquellos tipos de comportamiento que involucran conceptos. Por ejemplo, reconocer la constancia perceptiva, mediante la que se puede reconocer un objeto desde muchos puntos de vista diferentes; volver a identificar a los individuos a lo largo del tiempo; tener autocontrol anticipatorio (planificación); negociar motivos conflictivos, no solo programarlos; utilizar el método comparativo y el lenguaje. Estas críticas admiten que la robótica situada demuestra que el comportamiento libre de conceptos está más extendido de lo que creen muchos filósofos. No obstante, la lógica, el lenguaje y la acción humana requieren computación simbólica.

También muchos expertos en robótica rechazan los planteamientos más extremos de Brooks. El grupo de Alan Mackworth, uno de los muchos que trabajan en robots que juegan al fútbol, se atiene a la “deliberación reactiva”, que incluye la

percepción sensorial, la toma de decisiones en tiempo real, la planificación, el reconocimiento de plan, el aprendizaje y la coordinación.[7] La deliberación reactiva está buscando integrar la inteligencia artificial simbólica con la perspectiva situada. (Esto es: está construyendo sistemas híbridos: véase el capítulo IV).

En la robótica, en general, las representaciones son decisivas para el proceso de selección de las acciones, aunque menos para el de ejecución. Por eso, los bromistas que dijeron que las siglas de “IA” ahora significan “insectos artificiales” no acertaron del todo.

IA EVOLUTIVA

La mayoría asume que la IA requiere un diseño meticuloso. Dada la naturaleza implacable de los ordenadores, ¿cómo podría ser de otra manera? Bueno, pues sí puede.

Los robots evolutivos (entre los que se incluyen algunos robots situados), por ejemplo, son el resultado de combinar programación o ingeniería precisas con variaciones aleatorias. Evolucionan de manera impredecible, no diseñada al detalle.

La IA evolutiva en general tiene ese carácter, iniciado en la IA simbólica y utilizado también en el conexionismo. Entre sus muchas aplicaciones prácticas se encuentran el arte (un terreno que acoge bien lo impredecible) y el desarrollo de sistemas de seguridad críticos, como motores de aeronaves.

Un programa puede cambiarse a sí mismo (en vez de que lo reescriba un programador) e incluso puede mejorarse a sí mismo usando algoritmos genéticos (AG). Inspirados en la genética real, estos algoritmos permiten tanto la variación aleatoria como la selección no aleatoria. La selección necesita un criterio para medir el éxito o “función de aptitud” (equivalente a la selección natural en biología) colaborando con los AG. Definir la función de aptitud es crucial.

Un programa inicial orientado a la tarea de software evolutivo no puede resolver esa tarea de manera eficiente. Quizá no pueda resolverla en absoluto, ya que tal vez se trate de una recopilación incoherente de instrucciones o una red neuronal conectada al azar. Pero el programa general incluye algoritmos genéricos en segundo plano que pueden cambiar las reglas orientadas a la tarea. Los cambios, hechos al azar, se parecen a la mutación genética y al entrecruzamiento en biología. Así, en una instrucción programada se puede alterar un símbolo o se pueden "permutar" secuencias cortas de símbolos entre dos instrucciones.

Se comparan los diferentes programas de tareas dentro de una generación y los más eficaces se utilizan para engendrar la siguiente generación. Se pueden conservar también unos pocos más (elegidos al azar), para que no se pierdan las mutaciones potencialmente útiles que todavía no han tenido ningún efecto positivo. Conforme pasan las generaciones, aumenta la eficiencia del programa de tareas, y a veces se encuentra la solución óptima. (En algunos sistemas evolutivos se resuelve el problema de la asignación de mérito –véase el capítulo IV– con alguna variante del algoritmo "bucket-brigade" de John Holland, que identifica justo qué partes de un programa evolutivo complejo son responsables de su buen resultado).

Parte de la IA evolutiva es totalmente automática: el programa aplica la función de aptitud a cada generación y deja que evolucione sin supervisión. En este caso la tarea tiene que estar definida muy claramente (mediante la física de los motores de la aeronave, por ejemplo). El arte evolutivo, en cambio, suele ser muy interactivo (el artista selecciona lo mejor de cada generación), porque no se puede establecer claramente la función de aptitud (el criterio estético).

La mayor parte de la robótica evolutiva es interactiva intermitentemente. La anatomía del robot (por ejemplo, sensores y conexiones sensoriomotoras) y/o su controlador, el "cerebro") evolucionan automáticamente, pero como simulación. Para la mayoría de las generaciones no existe un robot físico, pero en la

generación número 500, quizá, el diseño evolucionado se pruebe en un dispositivo físico.

Las mutaciones inútiles tienden a no sobrevivir. Los investigadores de la universidad de Sussex descubrieron que uno de los dos “ojos” del robot y todos sus “bigotes” pueden perder las conexiones iniciales con la red neuronal controladora si la tarea no necesita ni visión profunda ni tacto.[8] (El córtex auditivo de los sordos congénitos o de los animales privados de entrada auditiva, de manera similar, se usa para la computación visual: el cerebro evoluciona a lo largo de la vida, no solo de una generación a otra).

La IA evolutiva puede producir grandes sorpresas. Por ejemplo, un robot situado que estaba desarrollando el mismo equipo de Sussex para que generase un movimiento hacia un objetivo evitando los obstáculos desarrolló un detector de orientación análogo a los que hay en el cerebro.[9] Uno de los elementos del mundo de ese robot era un triángulo de cartón blanco. Inesperadamente, en el controlador surgió una mini-red conectada al azar que reaccionaba a un gradiente claro / oscuro en una orientación particular (un lado del triángulo). Luego, evolucionaba como una parte integral de un mecanismo visuomotor, sus conexiones (en principio aleatorias) con las unidades motoras le permitieron al robot usar el objeto como ayuda para desplazarse. El mecanismo no funcionaba con un triángulo negro ni para el lado contrario. Y era un objeto independiente, ya que no había ningún sistema general de detectores de orientación. Era útil, no obstante. Este resultado asombroso era repetible en líneas generales. Usando redes neuronales de distintos tipos, el equipo de Sussex descubrió que todas las soluciones que funcionaban pasaban por haber desarrollado algún detector de orientación activo, así que la estrategia de comportamiento de alto nivel era la misma. (Los detalles exactos de la implementación variaban, pero solían ser muy parecidos).

En otra ocasión, el equipo de Sussex usó los algoritmos genéticos para diseñar circuitos eléctricos de hardware.[10] El objetivo

era desarrollar osciladores; sin embargo, lo que surgió fue un primitivo sensor de ondas de radio que captaba la señal de fondo del monitor de un ordenador cercano. Esto dependía de unos parámetros físicos imprevistos. Algunos eran predecibles (las propiedades de antena que tienen todas las placas de circuitos impresos), aunque el equipo no los había tenido en cuenta previamente, pero otros eran accidentales y aparentemente irrelevantes, como la proximidad espacial a un monitor de ordenador, el orden en que se habían colocado los interruptores analógicos y el hecho de que un soldador que estaba en un banco de trabajo cercano estuviera enchufado a la red. (Este resultado no se podía repetir: en la siguiente ocasión, a la antena de radio le podría afectar la composición química del papel de la pared).

El sensor de ondas de radio es interesante porque muchos biólogos (y filósofos) sostienen que no podía surgir nada radicalmente nuevo de la IA, ya que todos los resultados de un programa de ordenador (incluyendo los efectos aleatorios de los AG) deben estar comprendidos dentro del espacio de posibilidades definidos por él. Solo la evolución biológica, dicen, puede generar nuevos sensores perceptivos que permitan que un sensor visual débil de IA evolucione a uno mejor. Pero el primer sensor visual, dicen, podría surgir solo de un mundo físico gobernado por la causalidad. Una mutación genética al azar que provocara una sustancia química sensible a la luz podría introducir luz que ya estuviese presente en el mundo exterior dentro del medio del organismo. No obstante, y de manera similar, el inesperado sensor de ondas de radio traía ondas de radio al “medio” del dispositivo. Dependía en parte de la causalidad física (enchufes, etcétera), pero era un experimento de IA, no de biología.

La novedad radical en la IA sí requiere influencias exteriores, porque es cierto que un programa no puede superar su espacio de posibilidad, aunque estas influencias no tienen que ser físicas. Un sistema de algoritmos genéticos conectado a internet podría desarrollar novedades fundamentales al interactuar con un mundo virtual.

Otra sorpresa muy anterior de la IA evolutiva dio pie a una serie de investigaciones sobre la evolución propiamente dicha que siguen desarrollándose. El biólogo Thomas Ray utilizó algoritmos genéticos para simular la ecología de las selvas húmedas del trópico.[11] Observó la aparición espontánea de parásitos, la resistencia a estos y a superparásitos capaces de vencer esa resistencia. También descubrió que se pueden generar “saltos” bruscos en la evolución (fenotípica) mediante la sucesión de pequeñas mutaciones subyacentes (genotípicas). Los darwinianos ortodoxos ya creían esto, claro, pero es tan ilógico que algunos biólogos, como Stephen Jay Gould, afirman que debe de haber involucrados también procesos no darwinianos.

Hoy, se está variando y siguiendo la tasa de mutaciones simuladas sistemáticamente y los investigadores analizan “paisajes adaptativos”, “redes neutrales” [sic] y “derivas genéticas”. Este trabajo explica cómo se pueden preservar las mutaciones, aunque no hayan (todavía) aumentado la aptitud reproductiva, así que la IA ayuda a que los biólogos desarrollen la teoría evolutiva en general.

AUTOORGANIZACIÓN

El rasgo principal de los organismos biológicos es su capacidad de estructurarse a sí mismos. La autoorganización es el surgimiento espontáneo de orden desde un origen menos ordenado: una propiedad desconcertante, incluso casi paradójica, que no es evidente que pueda darse en las cosas inertes.

En líneas generales, la autoorganización es un fenómeno creativo. De la creatividad psicológica (tanto “histórica” como “individual”) se habló en el capítulo III y del aprendizaje asociativo autoorganizado (no supervisado) en el capítulo IV. Ahora nos interesan los tipos de autoorganización estudiados en la biología.

Entre los ejemplos se incluyen la evolución filogenética (una forma de creatividad histórica); la embriogénesis y la metamorfosis

(equivalentes a la creatividad individual en psicología); el desarrollo cerebral (la creatividad individual seguida por la creatividad histórica); y la formación celular (creatividad histórica cuando empezó la vida, creatividad individual después). ¿Cómo puede contribuir la IA a entenderlas?

Alan Turing explicó la autoorganización volviendo a lo esencial.[12] Se preguntó cómo algo homogéneo (por ejemplo, el óvulo indiferenciado) tenía la capacidad de originar estructuras. Se fijó en que la mayoría del desarrollo biológico añade un orden nuevo al orden anterior: así sucede en la secuencia de cambios en el tubo neural del embrión. Pero que el orden parta de la homogeneidad es el supuesto fundamental (y matemáticamente más sencillo).

Los embriólogos ya habían postulado la existencia de los "organizadores": sustancias químicas desconocidas que dirigían el desarrollo de forma desconocida. Turing tampoco pudo identificar los organizadores. En su lugar, planteó principios muy generales de la difusión química.

Demostró que, si se juntaban moléculas diferentes, los resultados dependían de las velocidades de reacción, sus concentraciones y las velocidades a las que sus interacciones destruirían / construirían moléculas. Lo hizo variando los números en ecuaciones químicas imaginarias e investigando los resultados. Algunas combinaciones numéricas producían solo mezclas de sustancias químicas sin forma, pero otras generaban orden; por ejemplo, picos de concentración regular de una molécula determinada. Esos picos químicos, dijo, podían expresarse biológicamente en forma de marcas superficiales (rayas) o de principios de estructuras repetidas, como pétalos o segmentos corporales. Los sistemas de reacción-difusión en tres dimensiones podrían producir un hueco, como hace el proceso de gastrulación en la etapa inicial del embrión.

Enseguida se vio que estas ideas eran tremendamente emocionantes. Resolvían el enigma antes insoluble de cómo surge

el orden a partir de un origen sin orden. Pero los biólogos de la década de 1950 no pudieron hacer mucho con ellas. Turing se había basado en el análisis matemático. Hizo a mano alguna simulación (tremendamente tediosa) y la pasó luego a un ordenador primitivo, pero su máquina no tenía bastante capacidad computacional para hacer las sumas pertinentes o para investigar las variaciones numéricas sistemáticamente, ni existían los gráficos por ordenador para convertir las listas de números en formas visualmente inteligibles.

Tanto la IA como la biología tuvieron que esperar cuarenta años para que se pudieran desarrollar las observaciones de Turing. El experto en gráficos por ordenador Greg Turk estudió las ecuaciones del mismo Turing, “congelando” los resultados de una ecuación antes de aplicar otra.[13] Este procedimiento, que recuerda al encendido / apagado de los genes, ilustraba el patrón desde otro patrón que había mencionado Turing, pero que no pudo analizar. En el modelo de IA de Turk, las ecuaciones de Turing generaban, además de las manchas de los dálmatas y rayas (como había pasado en las simulaciones que Turing hizo a mano), manchas de leopardo y de guepardo, retículas de jirafa y el camuflaje del pez león.

Otros investigadores utilizaron secuencias de ecuaciones más complicadas y obtuvieron, en consonancia, patrones más complejos. Algunos eran biólogos desarrollistas que ahora saben más sobre bioquímica real.

Por ejemplo, Brian Goodwin estudió el ciclo vital del alga Acetabularia.[14] Este organismo unicelular se transforma, a partir de un grumo informe, en un pedúnculo alargado; luego le crece una parte superior aplanada; después desarrolla un anillo de nudos alrededor del borde, que más tarde germinan un verticilo de laterales o ramas; finalmente, los laterales se funden y forman un sombrerete tipo paraguas. Los experimentos bioquímicos demuestran que más de treinta parámetros metabólicos están involucrados (por ejemplo, las concentraciones de calcio, la afinidad

entre el calcio y ciertas proteínas y la resistencia mecánica del citoesqueleto). El modelo computacional de Goodwin de la Acetabularia simulaba circuitos de retroalimentación complejos e iterativos en los que estos parámetros podían cambiar de un momento a otro. Resultaron diversas metamorfosis corporales.

Como Turing y Turk, Goodwin jugó con valores numéricos para ver cuáles generaban realmente formas nuevas. Usó solamente números dentro de los rangos observados en el organismo, aunque de manera aleatoria.

Descubrió que ciertos patrones, como la alternancia de concentraciones altas y bajas de calcio en la punta del pedúnculo (la simetría emergente de un verticilo), aparecían de manera recurrente. No dependían de la elección de un valor determinado de los parámetros, sino que surgían de forma espontánea si los valores establecidos entraban dentro de un rango concreto. Además, una vez originados los verticilos, subsistían. Así, según Goodwin, se podrían convertir en la base para las transformaciones que condujesen a otros rasgos frecuentes. Esto podía suceder tanto en la filogénesis como en la ontogénesis (creatividad histórica y también creatividad individual) de la evolución de uno de los miembros de un tetrápodo, por ejemplo.

Este modelo no generó nunca ningún paraguas que, posiblemente, requeriría parámetros adicionales que representasen interacciones químicas de la Acetabularia real hasta ahora desconocidas. O quizá esos paraguas sí que quedaban dentro del espacio de posibilidad del modelo y por tanto podrían surgir de él en principio, pero solo con unos valores numéricos limitados tan estrictamente que es improbable encontrarlos mediante una búsqueda aleatoria. (Tampoco se generaron los laterales, pero solo debido a la falta de potencia de cálculo de la máquina: habría que ejecutar el programa completo en un nivel más bajo para cada uno de los laterales).

Goodwin sacó una moraleja fascinante. Consideró que los verticilos eran formas “genéricas” que aparecían —a diferencia de los paraguas— en muchos animales y plantas. Esto sugiere que no son debidos a mecanismos bioquímicos muy específicos dirigidos por genes evolucionados de manera contingente, sino más bien a procesos generales (como la reacción-difusión) que se dan en la mayoría o incluso en todos los seres vivos. Estos procesos podrían componer la base de una biología “estructuralista”: una ciencia general de la morfología cuyas explicaciones serían anteriores a la selección darwiniana, aunque encajan en ella por completo. (Esta posibilidad estaba implícita en la reflexión de Turing y destacada por D’Arcy Thompson, un biólogo que él había citado, pero el propio Turing la ignoró).

La reacción-difusión está regulada por leyes fisicoquímicas que determinan las interacciones moleculares de cada punto, esto es, leyes representables en autómatas celulares. Cuando John von Neumann definió los autómatas celulares, indicó que en principio eran aplicables a la física. Los investigadores actuales de vida artificial utilizan los autómatas celulares para muchos propósitos y la generación de patrones biológicos es especialmente relevante. Por ejemplo, algunos autómatas celulares muy sencillos, definidos en una sola dimensión (una línea), pueden generar patrones extremadamente realistas, como los de las caracolas.[15]

Especialmente intrigante, quizá, sea el uso que hace la vida artificial de los autómatas celulares para intentar describir la “vida como podría ser”, no solo la “vida tal como la conocemos”.[16] Christopher Langton (que le dio nombre a la “vida artificial” en 1987) investigó numerosos autómatas celulares definidos al azar, observando su propensión a generar orden. Muchos producían solo caos, otros formaban estructuras aburridamente repetitivas o incluso estáticas, pero unos pocos generaron patrones sutilmente cambiantes, aunque relativamente estables, característicos, según Langton, de las cosas vivientes (y de la computación también). Sorprendentemente, estos autómatas celulares compartían el mismo valor numérico en un parámetro simple de la complejidad

informativa del sistema.[17] Langton sugirió que este “parámetro lambda” se aplica a todos los seres vivientes posibles, ya sea en la Tierra o en Marte.[18]

La autoorganización moldea órganos además de cuerpos completos. El cerebro, por ejemplo, se desarrolla mediante procesos evolutivos (a lo largo de la vida y de las generaciones) y también mediante aprendizaje no supervisado. Este aprendizaje puede tener resultados muy personales (históricamente creativos), pero el desarrollo cerebral temprano de cada individuo también crea estructuras neuronales predecibles. Por ejemplo, los monos recién nacidos tienen detectores de orientación que abarcan de forma metódica 360 grados y que no pueden haber aprendido a partir de su experiencia del mundo exterior, así que lo natural es asumir que vienen codificados en los genes. Pero no es así: surgen de manera espontánea de una red inicialmente aleatoria.

Esto se ha demostrado no solo mediante la simulación hecha por neurocientíficos en un ordenador biológicamente realista, sino también mediante IA “pura”. El investigador de IBM Ralph Linsker definió las redes multicapas unidireccionales (véase el capítulo IV) demostrando que en una actividad aleatoria (como el “ruido” dentro del cerebro embrionario) las reglas hebbianas simples son capaces de generar colecciones de detectores de orientación.[19]

Linsker no se basa solo en las demostraciones prácticas ni se concentra solo en los detectores de orientación: su teoría abstracta “infomax” se puede aplicar a cualquier red de este tipo. Esta establece que las conexiones de la red se desarrollan para maximizar la cantidad de información que se preserva cuando las señales se transforman en cada etapa del proceso. Todas las conexiones se forman bajo ciertas constricciones empíricas, como limitaciones bioquímicas y anatómicas. No obstante, las matemáticas garantizan que surja un sistema cooperativo de unidades comunicativas. La teoría infomax también se corresponde con la evolución filogenética. Hace que sea menos contrario a la lógica que, en la evolución de un sistema complejo, sea adaptiva

una sola mutación. La necesidad aparente de varias mutaciones simultáneas desaparece si cada nivel se puede adaptar de manera espontánea a una pequeña alteración de otro.

En lo que concierne a la autoorganización a nivel celular, se han hecho modelos tanto de la bioquímica intracelular como de la formación de las células / paredes celulares. Este trabajo aprovecha el de Turing sobre la reacción-difusión. No obstante, se basa más en conceptos biológicos que en ideas originadas en la vida artificial.

En suma, la IA ofrece muchas ideas teóricas acerca de la autoorganización, y tenemos artefactos autoorganizados por doquier.

VI

PERO ¿ES INTELIGENCIA DE VERDAD?

Supongamos que los sistemas futuros de IAF (pantallas o robots) igualasen los resultados humanos. ¿Tendrían inteligencia real, entendimiento real, creatividad real? ¿Tendrían identidad, integridad moral, libre albedrío? ¿Serían conscientes? Y sin conciencia, ¿podrían tener alguna de esas otras cualidades?

Estas no son preguntas científicas, sino filosóficas. Muchos sienten de manera intuitiva que la respuesta, en todos los casos, es “¡Obviamente, no!”.

Las cosas no son tan sencillas, sin embargo. Necesitamos argumentos meticulosos, no meras intuiciones sin analizar, pero esos argumentos demuestran que no hay respuestas irrefutables para esas preguntas. Esto se debe a que los conceptos en cuestión son de por sí muy controvertidos. Solo si los comprendiésemos todos satisfactoriamente, podríamos tener la seguridad de que esa IAF hipotética será, o no, de verdad inteligente. Dicho de otro modo: nadie lo sabe seguro.

Hay quien diría que no tiene importancia: lo importante es lo que realmente harán las IAF. No obstante, las respuestas podrían afectar a cómo nos relacionamos con ellas, como veremos.

Este capítulo, pues, no proporcionará respuestas inequívocas, pero propondrá que unas respuestas son más razonables que otras y mostrará cómo (algunos) filósofos han utilizado los conceptos de IA para ilustrar la naturaleza de la mente real.[1]

LA PRUEBA DE TURING

En un artículo publicado en la revista de filosofía Mind, Alan Turing describió la llamada prueba de Turing,[2] que plantea si alguien podría distinguir, durante el 30% del tiempo (en más de cinco minutos), si está interactuando con un ordenador o con una persona. En caso contrario, dio a entender, no habría ninguna razón para negar que un ordenador podía pensar de verdad.

Era irónico. Aunque aparecía en las primeras páginas, la prueba de Turing era un adjunto dentro de un artículo destinado en principio a ser un manifiesto para una IA futura. De hecho, Turing se lo describió a su amigo Robin Gandy como “propaganda” desenfadada que invitaba a la risa más que a la crítica seria.

No obstante, los filósofos saltaron. La mayoría sostuvo que, aunque las reacciones de un programa fuesen indistinguibles de las de un ser humano, eso no demostraría su inteligencia. La objeción más común fue (y sigue siendo) que la prueba de Turing se refiere solo al comportamiento observable, así que podría pasarla un zombi: algo que se comporta exactamente como nosotros, pero carece de conciencia.

Esta objeción asume que la inteligencia requiere conciencia y que los zombis son lógicamente posibles. Veremos (en la sección “IA y conciencia fenoménica”) que algunas versiones de conciencia implican que el concepto de zombi es incoherente. Si tienen razón, entonces ninguna IAF podría ser un zombi. A ese respecto, la prueba de Turing estaría justificada.

La prueba de Turing les interesa enormemente a los filósofos (y al público en general), pero no ha sido importante para la IA. Por lo general, el objetivo de la IA es proporcionar herramientas útiles, no imitar la inteligencia humana y, menos aún, hacerle creer a los usuarios que están interactuando con una persona. Es verdad que los investigadores de IA deseosos de publicidad a veces aseguran y/o permiten que los periodistas aseguren que su sistema pasa la prueba de Turing; no obstante, esas pruebas no concuerdan con la descripción de Turing. Por ejemplo, el modelo de Ken Colby, PARRY, “engañó” a los psiquiatras para que pensaran que estaban leyendo entrevistas con paranoicos, porque naturalmente asumieron que se trataba de pacientes

humanos.[3] De igual manera, el arte digital se suele atribuir a un ser humano si no hay indicios de que pueda haber participado una máquina.

Lo más parecido a una prueba de Turing genuina es la competición anual de Loebner (que ahora se celebra en Bletchley Park). Las normas actuales prescriben interacciones de veinticinco minutos mediante veinte preguntas seleccionadas previamente para poner a prueba la memoria, el razonamiento, los conocimientos generales y la personalidad. Los jueces tienen en cuenta la relevancia, la corrección y la claridad, y plausibilidad de la expresión y la gramática. Hasta el momento, ningún programa ha engañado a los jueces de Loebner durante el 30% del tiempo. (En 2014, un programa que decía ser un niño ucraniano de trece años engañó al 33% de sus interrogadores, pero a los hablantes no nativos se les perdonan fácilmente los errores, en especial a los niños).

LOS NUMEROSOS PROBLEMAS DE LA CONCIENCIA

No existe algo que sea “el” problema de la conciencia: en realidad son muchos. La palabra “conciencia” se utiliza para hacer muchas distinciones diferentes; deliberada / espontánea; con / sin atención; accesible / inaccesible; declarativa / no declarativa; reflexiva / irreflexiva, etcétera. No hay una única explicación que las aclare todas.

Los contrastes que se han enumerado son funcionales. Muchos filósofos admitirían que en principio se podría entender la conciencia en términos de procesamiento de la información y/o neurocientíficos.

Pero la conciencia fenoménica (sensaciones –como la melancolía o el dolor– o “qualia” –el término técnico que usan los filósofos–) parece ser diferente. La misma existencia de los qualia en un universo básicamente material es un enigma metafísico muy conocido.

David Chalmers lo llama “el problema difícil”[4] y dice que es ineludible: “[Debemos] tomarnos en serio la conciencia… Volver a definir el problema como la explicación de cómo se representan algunas funciones cognitivas o de comportamiento es inaceptable”.

Se han sugerido diversas soluciones muy especulativas, incluyendo la versión del mismo Chalmers del pampsiquismo; una teoría, según su

propia confesión, “atroz o incluso descabellada” según la cual la conciencia fenoménica es una propiedad irreductible del universo, análoga a la masa o a la carga eléctrica. Otros teóricos han recurrido a la física cuántica usando un misterio para resolver otro, según dicen sus oponentes. Colin McGinn ha llegado a afirmar que los seres humanos son incapaces por su constitución de comprender la relación causal entre cerebro y qualia, igual que los perros no pueden comprender la aritmética.[5] Y Jerry Fodor, un destacado filósofo de la ciencia cognitiva, cree que “nadie tiene la menor idea de cómo podría ser consciente algo material. Ni siquiera sabemos cómo sería tener la menor idea de cómo podría ser consciente algo material”.[6]

En suma, muy pocos filósofos afirman comprender la conciencia fenoménica y a los que sí lo hacen no los cree casi nadie. Este tema es una ciénaga filosófica.

CONCIENCIA ARTIFICIAL

Los pensadores simpatizantes de la IA se plantean la conciencia de dos formas. Una pasa por construir modelos de conciencia computerizados: esto se llama “máquina consciente” (MC). La otra (característica de los filósofos influidos por la IA) es analizarla en términos principalmente computacionales sin hacer modelos.

Una IA fuerte verdaderamente inteligente poseería conciencia funcional. Por ejemplo, concentraría su mirada y su atención en cosas distintas en momentos distintos. Un sistema con nivel humano podría también deliberar y reflexionar sobre sí mismo. Podría generar ideas creativas y hasta evaluarlas voluntariamente. Sin esas capacidades, no podría comportarse de forma aparentemente inteligente.

La conciencia fenoménica puede estar involucrada cuando un ser humano evalúa ideas creativas (véase el capítulo III). De hecho, muchos dirían que está presente en todas las diferencias “funcionales”. Sin embargo, los investigadores de la conciencia artificial (que estudian todos la conciencia funcional) por lo general ignoran la conciencia fenoménica. (Unos cuantos valientes –¿temerarios?– aseguran que su sistema de IA ya la tiene “a su manera” porque basa las discriminaciones en entradas perceptuales, por ejemplo, la luz. Que eso

suponga la presencia de experiencia visual es –para decirlo con suavidad– muy dudoso).

Un proyecto interesante de conciencia artificial es LIDA (Learning Intelligent Distribution Agent), creado en Memphis por el grupo de Stan Franklin.[7] El nombre LIDA designa dos cosas. Una es un modelo conceptual (una teoría computacional expresada verbalmente) de conciencia (funcional). La otra es una implementación parcial y simplificada de ese modelo teórico.

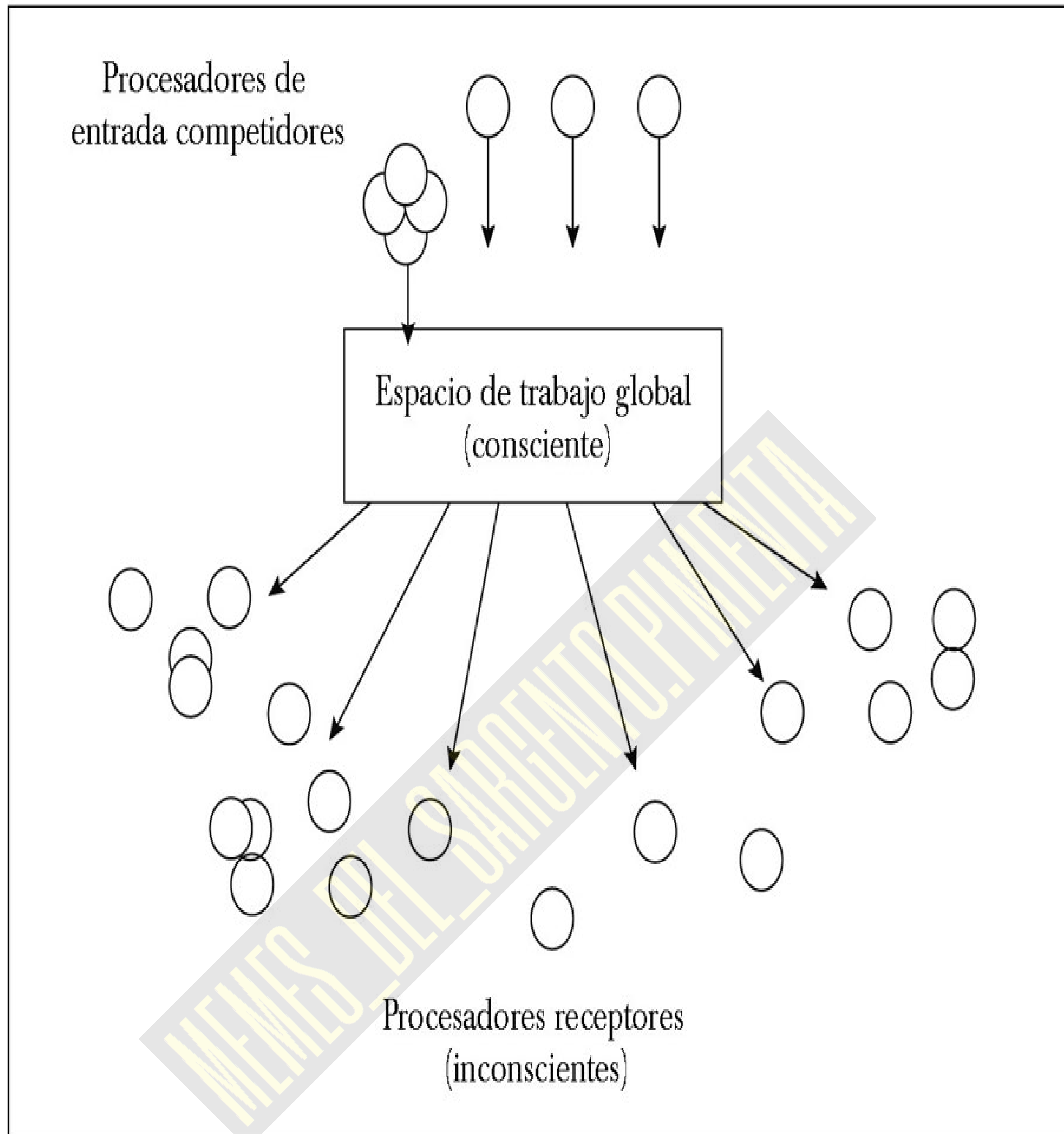
Ambas se utilizan con propósitos científicos (el objetivo principal de Franklin), pero la segunda también tiene aplicaciones prácticas. LIDA se puede adaptar a dominios de problemas específicos, por ejemplo, a la medicina.

A diferencia de SOAR, ACT-R y CYC (véase el capítulo II), LIDA es muy reciente. La primera versión (construida para que la marina de Estados Unidos organizase los nuevos trabajos para los marineros que terminaban su periodo de servicio) salió en 2011. La versión actual se ocupa de la atención y sus efectos sobre el aprendizaje en varios tipos de memoria (episódica, semántica y procedimental); y el control sensoriomotor se está implementando ahora en la robótica, pero siguen faltando muchos elementos, entre ellos el lenguaje. (La descripción posterior concierne al modelo conceptual, al margen de qué aspectos se han implementado ya).

LIDA es un sistema híbrido que incorpora disfusión de la activación, representaciones dispersas (véase el capítulo IV) y programación simbólica. Está basado en la teoría de conciencia neuropsicológica de Bernard Baars, la teoría del espacio de trabajo global (Global Workspace Theory o GWT por sus siglas en inglés).[8]

La GWT considera que el cerebro es un sistema distribuido (véase el capítulo II) en el que compiten una gran cantidad de subsistemas especializados funcionando en paralelo por el acceso a la memoria de trabajo (véase el Cuadro 2, en la página siguiente). Los elementos aparecen por orden (la corriente de la conciencia) pero se “retransmiten” a todas las áreas corticales.

Si un elemento retransmitido, derivado desde un órgano de los sentidos o de otro subsistema, desencadena una reacción en un área determinada, esa reacción puede ser lo bastante fuerte como para ganar la competición por la atención que controla de manera activa el acceso a la conciencia. (Las percepciones / representaciones novedosas tienden a captar la atención, mientras que los elementos repetidos se desvanecen de la conciencia). Los subsistemas son complejos por lo general. Algunos están anidados jerárquicamente y muchos comparten conexiones asociativas de tipos diversos. Una mezcla de contextos inconscientes (organizados en memorias diferentes) conforman la experiencia consciente, tanto evocando como modificando los elementos del espacio de trabajo global. Los contenidos de la atención, por orden, adaptan los contextos que persisten y crean conocimiento de varios tipos.



Cuadro 2. Un espacio de trabajo global en un sistema distribuido. El sistema nervioso incluye diversos procesadores inconscientes especializados (analizadores perceptuales, sistemas de salida, sistemas de planificación, etcétera). La interacción, coordinación y control de estos especialistas inconscientes requiere un intercambio de información central o “espacio de trabajo global”. Los especialistas en entradas pueden cooperar y competir por el acceso a él. En el caso que aquí se muestra, cuatro procesadores de entrada cooperan para ubicar un mensaje global, que luego se transmite al sistema como un todo.

Adaptado de la p. 88 de B. J. Baars, A Cognitive Theory of Consciousness (Cambridge, Cambridge University Press, 1988), reproducido con permiso.

Estos contenidos, cuando se retransmiten, guían la elección de la siguiente acción. Muchas acciones son cognitivas, como construir o adaptar las representaciones internas. Las reglas morales se almacenan (en la memoria semántica) como procedimientos para evaluar acciones potenciales. También las reacciones percibidas / previstas de otros agentes sociales pueden influir en las decisiones.

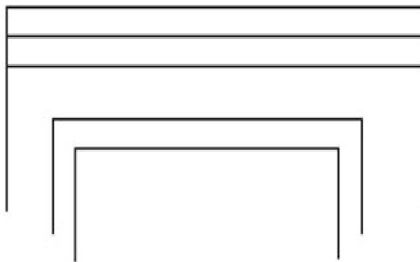
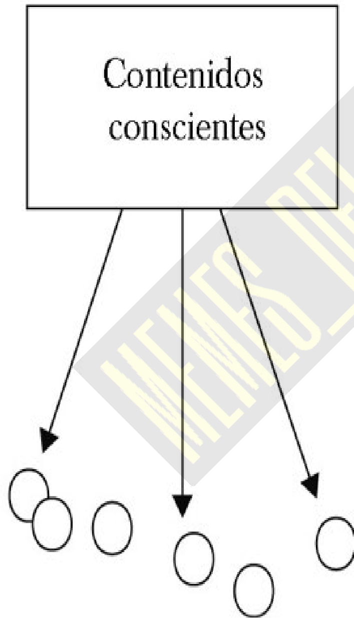
En la planificación, por ejemplo (véase el capítulo II), las intenciones se representan como estructuras en su mayor parte inconscientes, pero relativamente de alto nivel, que pueden llevar a imágenes objetivo conscientes (seleccionadas por elementos destacados de la percepción, la memoria o la imaginación). Estas reclutan subobjetivos relevantes; "reclutan" más que "recuperan", ya que son los mismos subobjetivos los que deciden la relevancia. Como todos los subsistemas corticales, están al acecho esperando a que los desencadene algún elemento emitido, en este caso, una imagen objetivo apropiada. LIDA puede transformar un esquema de acción orientado al objetivo seleccionado en acciones motoras ejecutables de bajo nivel que reaccionen a elementos concretos de un medio impredecible y cambiante.

La teoría de Baars (y la versión de Franklin) no se inventó en el taller de un científico informático. Al contrario, se diseño para tener en cuenta gran variedad de fenómenos psicológicos conocidos y una amplia gama de pruebas experimentales (véase el Cuadro 3, en la página siguiente), pero estos autores afirman que también soluciona algunos rompecabezas psicológicos antes sin resolver.

Por ejemplo, dicen que GWT / LIDA resuelve el muy controvertido problema de la "integración", que trata de cuántas entradas de diferentes sentidos en diferentes áreas del cerebro (por ejemplo, el tacto, la apariencia, el olor y el sonido de un gato) se le atribuyen a una única y misma cosa. Según Franklin y Baars, también explica cómo evita la mente humana el problema del marco (véase el capítulo II). Al generar analogías creativas, por ejemplo, no hay una ejecutiva central que busque la estructura completa de los datos para los objetos relevantes.

Más bien, si un subsistema reconoce que algún objeto transmitido casa o se aproxima a lo que (siempre) está buscando, compite por entrar al espacio de trabajo global.

Franklin usa LIDA para investigar teorías de psicología cognitiva y neurociencia integrando una evidencia experimental muy diversa. Por ejemplo, ha simulado el “parpadeo de la atención”, en el que el sujeto no consigue detectar un segundo estímulo visual que se le presenta poco después del primero. Hay otras teorías y modelos computacionales del parpadeo de la atención, pero la mayoría están diseñados para responder a preguntas aisladas. El modelo de Franklin surge de una teoría unificada de la cognición de sistemas y niveles. (Existe otro modelo unificado del parpadeo de la atención basado en ACT-R, pero ACT-R no incluye procesamiento emocional o visión de alto nivel, así que no puede explicar todos los resultados experimentales).

Teoría del espacio global de trabajo	Equivalentes
Contextos 	Conjunto de expectativas (Bruner) Disposiciones perdurables e intenciones momentáneas (Kahneman) Memoria activada (Bransford) Esquemas activados (Norman & Rumelhart) Sistema de acción dominante (Shallice) Aufgabe (Würzburg School, Ach)
	Conciencia Atención Capacidad central limitada Memoria a corto plazo Memoria de trabajo (Baddeley J. Anderson) Procesos estratégicos/controlados (Shiffrin & Schneider)
Procesadores especializados inconscientes	Procesos automáticos Memoria a largo plazo Facultades Especializaciones adaptivas (Rozin) Módulos (Fodor) Procesadores de distribución paralela (Rumelhart & McClelland)

Cuadro 3. Semejanzas entre los términos del espacio de trabajo global y otros conceptos muy extendidos. Cada una de estas ideas conocidas se define (mediante GWT) en términos de funcionamiento inconsciente y consciente.

Adaptado de la p. 44 de B. J. Baars, A Cognitive Theory of Consciousness, Cambridge, Cambridge University Press, 1988. Reproducido con permiso.

Este enfoque de la IA recuerda a los "demonios" de Pandemónium y a las arquitecturas de pizarra utilizadas para implementar los sistemas de producción (véase el capítulos I y II). No es sorprendente que esas ideas, ya que inspiraron la teoría neuropsicológica de Baars, finalmente llevasen a LIDA. La rueda teórica ha dado una vuelta completa.[9]

IA Y CONCIENCIA FENOMÉNICA

Los profesionales de la conciencia artificial ignoran el problema "difícil", pero hay tres filósofos que basándose en la IA la han abordado de frente: Paul Churchland, Daniel Dennett y Aaron Sloman. Decir que sus respuestas son controvertidas es quedarse corto. En lo que concierne a la conciencia fenoménica, no obstante, eso es parte del proceso.

El "materialismo eliminativo" de Churchland niega la existencia de pensamientos y experiencias inmateriales,[10] y los identifica con estados cerebrales.

Su teoría científica (en parte computacional –conexionismo–, en parte neurológica) define un "espacio de degustación" de cuatro dimensiones, que localiza y representa metódicamente las discriminaciones subjetivas (qualia) dentro de estructuras neuronales específicas. Las cuatro dimensiones reflejan los cuatro tipos de receptores de gusto que hay en la lengua.

Para Churchland esta no es una cuestión de correlaciones entre mente y cerebro: tener una experiencia de gusto significa simplemente que el cerebro visita un punto en particular del espacio sensorial definido de manera abstracta. Esto implica que toda conciencia fenoménica consiste únicamente en que el cerebro está en un punto concreto dentro

de una especie de hiperespacio localizable empíricamente. En ese caso, ningún ordenador (con la posible excepción de una emulación del cerebro completo) podría tener conciencia fenoménica.

Dennett también niega la existencia de experiencias ontológicamente distintivas más allá de los acontecimientos físicos. (Por eso una reacción común a su provocativo libro es: “No la conciencia explicada, sino desechada”).[11]

Experimentar, según él, es discriminar. Pero discriminar algo que existe en el mundo material no significa que se le da vida a otra cosa en otro mundo inmaterial. Esto lo expresa en una conversación imaginaria:

[OTTO:] Me parece que has negado la existencia de los fenómenos más indudablemente reales que hay: los claros y distintos que ni siquiera Descartes en sus Meditaciones pudo negar.

[DENNETT:] En un sentido, tienes razón, eso es lo que estoy negando que existe. Pensemos en el fenómeno de la difusión del color neón. Parece haber un círculo rosa brillante en la cubierta.

[Está describiendo una ilusión óptica causada por líneas rojas y negras sobre un papel blanco brillante.]

[OTTO:] Pues sí que parece.

[DENNETT:] Pero no hay ningún círculo rosa de verdad.

[OTTO:] Cierto, pero ¡sí que parece haberlo!

[DENNETT:] Cierto.

[OTTO:] Entonces, ¿dónde está?

[DENNETT:] ¿Dónde está qué?

[OTTO:] El círculo rosa brillante.

[DENNETT:] No hay ninguno. Creía que acababas de reconocerlo.

[OTTO:] Bueno, sí, no hay ningún círculo rosa brillante en la página, pero sí que parece haber uno.

[DENNETT:] Cierto. Parece que hay un círculo rosa brillante.

[OTTO:] Entonces hablemos de ese círculo.

[DENNETT:] ¿De cuál?

[OTTO:] Del que parece que hay.

[DENNETT:] No hay tal cosa como un círculo rosa que solo parece estar.

[OTTO:] Mira, no solo digo que parece haber un círculo rosa brillante; ¡es verdad que parece haber un círculo rosa brillante!

[DENNETT:] Me he apresurado al darte la razón… Hablas en serio cuando dices que parece haber un círculo rosa brillante.

[OTTO:] Mira, no solo lo digo en serio, es que no solo pienso que parece haber un círculo rosa brillante, es que ¡de verdad parece haber un círculo rosa brillante!

[DENNETT:] Menuda la has hecho. Has caído en una trampa, junto con muchos otros. Pareces creer que existe una diferencia entre pensar (juzgar, decidir, tener la convicción de) que algo te parece rosa y algo que realmente parece rosa, pero no hay diferencia. No existe ese fenómeno de parecer de verdad más allá del fenómeno de juzgar de una manera u otra que algo es así.

Dicho de otro modo, no se puede satisfacer la petición de explicar los qualia.

No existen tales cosas. Aaron Sloman no está de acuerdo. Reconoce la existencia real de los qualia, pero lo hace de forma poco habitual: los analiza como aspectos de la máquina virtual multidimensional que llamamos mente (que veremos en la siguiente sección).

Los qualia, dice, son estados computacionales internos.[12] Tienen efectos causales sobre el comportamiento (por ejemplo, expresiones faciales involuntarias) y/o sobre otros aspectos del procesamiento de la información de la mente. Pueden existir solo en máquinas virtuales de complejidad estructural significativa (destaca los tipos de recursos computacionales reflexivos necesarios). Se puede acceder a ellos a través de otras partes de la máquina virtual en cuestión y no necesariamente se expresan en la conducta (de ahí su privacidad). Además, no siempre se pueden describir en términos verbales mediante niveles superiores de la mente que se supervisan a sí mismos (de ahí su inefabilidad).

Esto no quiere decir que Sloman identifique los qualia con los procesos cerebrales (como Churchland), ya que los estados computacionales son aspectos de las máquinas virtuales: no se pueden definir con el lenguaje de las descripciones físicas, pero pueden existir y tener efectos causales solo cuando se los implementa en algún mecanismo físico subyacente.

¿Qué pasa con la prueba de Turing? Tanto el análisis de Dennett como el de Sloman implican (y Dennett lo asegura explícitamente) que los zombis son imposibles.[13] Esto es porque, para ellos, el concepto de zombi es incoherente. Si se da el comportamiento y/o máquina virtual adecuados, la conciencia (para Sloman, incluso incluyendo los qualia) está garantizada. La prueba de Turing, por tanto, se salva de la objeción de que podría "pasarla" un zombi.

¿Y qué pasa con una hipotética IAF? Si Dennett tiene razón, tendría toda la conciencia que tenemos nosotros, que no incluiría a los qualia. Si Sloman tiene razón, tendría conciencia fenoménica igual que nosotros.

LAS MÁQUINAS VIRTUALES Y EL PROBLEMA MENTE-CUERPO

El "funcionalismo" de Hilary Putnam de la década de 1960 utilizaba la noción de máquinas de Turing y la distinción (entonces novedosa) entre software / hardware para afirmar que, en efecto, la mente es lo que el cerebro hace.[14]

La división metafísica (cartesiana) entre dos sustancias completamente diferentes dio lugar a la división conceptual entre niveles de descripción. La analogía de programa versus ordenador admitía que la “mente” y el “cuerpo” son de hecho muy diferentes, pero era del todo compatible con el materialismo. (Se discutió y se sigue discutiendo acaloradamente si podía incluir los qualia).[15]

Aunque en la década de 1960 existían varios programas de IA misteriosos (véase el capítulo I), los filósofos funcionalistas rara vez pensaron en ejemplos específicos. Se centraban en los principios generales, como la computación de Turing. Solo cuando a mediados de la década de 1980 surgieron los PDP (véase el capítulo IV), muchos filósofos se preguntaron cómo funcionaban realmente los sistemas de IA. Incluso entonces, muy pocos preguntaron qué funciones computacionales concretas hacían posible (por ejemplo) el razonamiento o la creatividad.

La mejor manera de comprender estos asuntos es adoptar el concepto de máquinas virtuales en el ámbito científico informático. En vez de decir que la mente es lo que el cerebro hace, se debería decir (según Sloman) que la mente es la máquina virtual (o, más bien, el conjunto integrado de muchas máquinas virtuales diferentes) implementada en el cerebro. (Sin embargo, la postura de que la mente es una máquina virtual tiene una implicación muy contraria a la razón: véase la sección “¿Es esencial la neuroproteína?”).

Como se ha explicado en el capítulo I, las máquinas virtuales son reales y tienen efectos causales reales: no hay ninguna interacción metafísica y misteriosa entre la mente y el cerebro. Así que la importancia filosófica de LIDA, por ejemplo, radica en que establece un conjunto organizado de máquinas virtuales que demuestra cómo son posibles los diversos aspectos de la conciencia (funcional).

El enfoque de la máquina virtual corrige un aspecto primordial del funcionalismo, la hipótesis de los Sistemas de Símbolos Físicos (physical symbol system, PSS por sus siglas en inglés).[16] En la década de 1970, Allen Newell y Herbert Simon definieron el PSS como “un conjunto de entidades llamadas símbolos que son patrones físicos que pueden aparecer como componentes de otro tipo de entidad llamada

expresión (o estructura simbólica) […] [Dentro] de una estructura simbólica, […] los ejemplos (o muestras) de símbolos se relacionan de alguna manera física (por ejemplo, que una muestra esté al lado de otra)”. Los procesos existen, según ellos, para crear y modificar estructuras de símbolos, como podrían ser los procesos definidos mediante IA simbólica. Y añadieron que “un PSS tiene los medios necesarios y suficientes para la acción general inteligente”. Dicho de otro modo, la mente-cerebro es un PSS.

Desde la perspectiva de que la mente es una máquina virtual, la hipótesis debería haberse llamado hipótesis de los sistemas de símbolos implementados físicamente (a esto no le vamos a dar un nombre con siglas), ya que los símbolos son contenidos de las máquinas virtuales, no máquinas físicas.

Esto implica que el tejido neuronal no es necesario para la inteligencia, a menos que sea el único sustrato material en el que se pueden implementar las máquinas virtuales en cuestión.

La hipótesis de los sistemas de símbolos físicos (y de la mayoría de la IA primera) asumía que una representación o símbolo físico es un elemento de una máquina / cerebro que se puede aislar y localizar con precisión. El conexionismo daría una interpretación muy diferente de las representaciones (véase el capítulo IV). Las consideraba redes completas de células, no neuronas claramente localizables, y concebía los conceptos como constricciones en parte conflictivas, no como definiciones lógicas estrictas. Esto fue muy atractivo para los filósofos que conocían el concepto de “aires de familia” de Ludwig Wittgenstein.[17]

Más tarde, los que trabajaban en robótica situada negaron que el cerebro contuviese ninguna representación (véase el capítulo V). Algunos filósofos adoptaron esta postura, pero David Kirsh, por ejemplo, sostuvo que se necesitan representaciones constitutivas (y la computación simbólica) para todo comportamiento que requiera conceptos, incluidos la lógica, el lenguaje y la acción deliberativa.[18]

SIGNIFICADO Y COMPRENSIÓN

Según Newell y Simon, cualquier PSS que lleva a cabo los cálculos correctos es realmente inteligente. Tiene "los medios necesarios y suficientes para actuar con inteligencia". El filósofo John Searle llamó a esta afirmación "IA fuerte".[19] (La "IA débil" solo estipulaba que los modelos de IA podían ayudar a los psicólogos a formular teorías coherentes).

Searle sostuvo que la IA fuerte estaba equivocada. Puede que haya computación simbólica dentro de la cabeza (aunque él lo dudaba), pero eso solo no puede proporcionar inteligencia. Para ser más exactos, no puede proporcionar "intencionalidad", el término técnico que le dan los filósofos al significado o la comprensión.

Searle se basó en un experimento difícil que sigue siendo controvertido: "Searle está en una habitación sin ventanas, con una ranura por la que pasan hojas de papel con 'garabatos' y 'garabatas'. Hay una caja de hojas con dibujos similares y un libro de reglas que dice que, si un entra un garabato, entonces Searle debe sacar por la ranura un bongo bongo, o quizá pasar por una larga secuencia de parejas de garabatos antes de pasar una hoja por la ranura. Sin que Searle lo sepa, los dibujos están en escritura china, el libro de normas es un programa chino de procesamiento de lenguajes naturales y las personas chinas que están fuera de la habitación lo están usando para responder sus preguntas. No obstante, Searle entró en la habitación incapaz de entender chino y seguirá sin entenderlo cuando salga. Conclusión: El cálculo formal solo (que es lo que está haciendo Searle en la habitación) no puede generar intencionalidad, así que la IA fuerte está equivocada y es imposible que los programas de IA tengan una comprensión genuina". (Este argumento, llamado "la habitación china", estaba destinado en principio a la IA simbólica, pero luego se generalizó y se aplicó al conexionismo y a la robótica).

La reivindicación de Searle era que el "significado" atribuido a los programas de IA proviene enteramente de los usuarios / programadores humanos. Son arbitrarios respecto al programa mismo, que semánticamente está vacío. Al ser "todo sintaxis y ninguna semántica", el mismo programa podría ser igual de interpretable que una calculadora de impuestos o una coreografía.

En algunos casos es cierto, pero hay que recordar que Franklin afirma que los modelos de LIDA situaban, incluso encarnaban la cognición mediante el emparejamiento estructurado entre los sentidos, los activadores y el medio. Recuerde el lector también el circuito de control que evolucionó hasta convertirse en un detector de orientación robótico (véase el capítulo V). Que se llame “detector de orientación” no es arbitrario. Su existencia misma depende de su evolución como detector de orientación y de que sea útil para que el robot alcance su objetivo.

El último ejemplo es relevante, y no solo porque algunos filósofos consideran que la evolución es el origen de la intencionalidad. Ruth Millikan, por ejemplo, sostiene que el pensamiento y el lenguaje son fenómenos biológicos, cuyos significados dependen de nuestra historia evolutiva.[20] Si eso fuese cierto, ninguna IAF podría tener entendimiento verdadero.

Otros filósofos de la ciencia (como los mismos Newell y Simon) definen la intencionalidad en términos causales, pero tienen dificultades para representar las aseveraciones no verídicas: si alguien dice ver una vaca, pero allí no hay ninguna vaca que cause esas palabras, ¿cómo pueden significar vaca?

En suma, ninguna teoría de la intencionalidad satisface a todos los filósofos. Que la inteligencia genuina requiera comprensión es otra razón por la que nadie sabe si nuestra hipotética IA fuerte sería realmente inteligente.

¿ES ESENCIAL LA NEUROPROTEÉNA?

Una de las razones de Searle para rechazar la IA fuerte fue que los ordenadores no están hechos de neuroproteína. Dijo que la intencionalidad la genera la neuroproteína tanto como la clorofila causa la fotosíntesis. Puede que la neuroproteína no sea la única sustancia del universo capaz de sustentar la intencionalidad y la conciencia, pero es obvio que el metal y el silicio son incapaces de hacerlo.

Esto es ir demasiado lejos. Es cierto que sugerir que unos ordenadores de lata podrían experimentar melancolía o dolor o entender

verdaderamente el lenguaje va en contra del sentido común, pero que la neuroproteína genere qualia no es menos ilógico ni menos problemático para la filosofía. (Sin embargo, algo que va en contra del sentido común puede ser cierto).

Si aceptamos el análisis de los qualia de la máquina virtual de Sloman, esta dificultad concreta se desvanece. No obstante, la concepción de la mente en su conjunto como máquina virtual conlleva otra dificultad similar. Si se implementase en hardware con IA una máquina virtual que cumpliese los requisitos de mente, entonces esa misma mente existiría en la máquina o quizá en varias máquinas. Así que una mente considerada como máquina virtual implica la posibilidad, en principio, de una inmortalidad personal (clonada) computerizada. Para la mayoría (no obstante, véase el capítulo VII), esto no es menos contraintuitivo que los ordenadores conteniendo qualia.

Si la neuroproteína es realmente la única sustancia capaz de sustentar máquinas virtuales a escala humana, podemos rechazar la propuesta de la “inmortalidad clonada”. Pero ¿lo es? No lo sabemos.

Quizá las propiedades que le permiten a la neuroproteína aplicar el amplio rango de cálculos que la mente lleva a cabo sean especiales, quizá muy abstractas. Por ejemplo, debe ser capaz de crear (con bastante rapidez) moléculas estables (y que se puedan almacenar) y que aun así también sean flexibles. Debe ser capaz de formar estructuras y conexiones entre las estructuras que tienen propiedades electroquímicas que les permiten transmitirse información. Es posible que otras sustancias, en otros planetas, puedan hacer también estas cosas.

NO SOLO EL CEREBRO, EL CUERPO TAMBIÉN

Algunos filósofos de la mente sostienen que el cerebro recibe demasiada atención. El cuerpo completo, dicen, es un mejor encuadre.[21]

Su postura se basa en la fenomenología continental,[22] que destaca la “forma de vida humana”. Esta abarca la conciencia significativa (que incluye los “intereses” humanos en los que se fundamenta nuestro sentido de la relevancia) y la encarnación del agente.

La encarnación o embodiment es vivir en un cuerpo vivo situado y partícipe activo de un medio dinámico. El medio (y la participación) es tanto físico como sociocultural. Las propiedades psicológicas principales no son el razonamiento ni el pensamiento, sino la adaptación y la comunicación.

Los filósofos de la encarnación no pierden el tiempo con la IA simbólica y la consideran excesivamente cerebral; solo les conceden importancia a los enfoques basados en la cibernética (véanse los capítulos I y V) y como, desde ese punto de vista, la inteligencia genuina se basa en el cuerpo, ninguna IA fuerte en pantalla podría ser realmente inteligente. Incluso si el sistema en pantalla es un agente autónomo acoplado estructuralmente a un medio físico, no contaría (a pesar de Franklin)[23] como encarnado.

¿Qué pasa con los robots? Al fin y al cabo, los robots son seres físicos anclados y adaptados al mundo real. De hecho, estos filósofos a veces encomian la robótica situada, pero ¿tienen cuerpos los robots? ¿O intereses? ¿O formas de vida? ¿Están vivos siquiera?

Los fenomenólogos dirían "¡Por supuesto que no!". Podrían citar la famosa observación de Wittgenstein: "Si un león pudiese hablar, no lo entenderíamos". La forma de vida del león es tan diferente de la nuestra que la comunicación sería casi imposible. Bien es cierto que la psicología de un león y la nuestra se solapan lo suficiente (por ejemplo, el hambre, el miedo, el cansancio, etcétera) para que alguna comprensión y empatía mínimas pudieran ser factibles, pero al "comunicarse" con un robot no se dispondría siquiera de eso. (Por eso preocupa tanto la investigación sobre robots acompañantes: véanse los capítulos III y VII).

COMUNIDAD ÉTICA

¿Aceptaríamos (¿deberíamos aceptar?) una IAF de nivel humano como miembro de nuestra comunidad ética? De ser así, habría consecuencias prácticas significativas, ya que afectaría a la interacción entre humanos y ordenadores de tres maneras.

Primero, la IAF sería objeto de nuestra inquietud ética, igual que los animales. Respetaríamos sus intereses hasta cierto punto. Si una IA fuerte nos pidiese que interrumpiéramos nuestro descanso o el crucigrama que estamos haciendo para ayudarle a conseguir un objetivo de “suma prioridad”, lo haríamos. (¿No se ha levantado nunca el lector de su sillón para pasear al perro o dejar salir al jardín a un insecto?). Cuanto más importantes creamos que son sus intereses para la IAF, más obligados nos sentiríamos a respetarlos. Sin embargo, ese juicio dependería en gran parte de si le atribuyésemos a la IAF una conciencia fenoménica (que incluyese emociones profundas).

Segundo, consideraríamos que sus acciones pueden ser valoradas moralmente. Los drones asesinos actuales no tienen responsabilidad moral (a diferencia de sus usuarios / diseñadores: véase el capítulo VII), pero ¿quizá una IAF verdaderamente inteligente la tendría? Presumiblemente, sus decisiones se podrían ver afectadas por nuestra reacción, por nuestro elogio o nuestra reprimenda, si no, no habría comunidad. La IA podría aprender a ser “moral” igual que un niño pequeño (o un perro) aprenden a comportarse bien o un niño un poco mayor aprende a ser amable. (La amabilidad requiere del desarrollo de lo que los psicólogos cognitivos llaman “teoría de la mente”, que interpreta el comportamiento de la gente en términos de voluntad, intención y creencia). Hasta el castigo puede estar justificado por motivos instrumentales.

Y tercero, lo convertiríamos en objeto de discusión y persuasión sobre decisiones morales. La IA podría hasta ofrecer consejo moral a los humanos. Para entablar en serio esta conversación, habría que tener la seguridad de que (además de tener inteligencia como la humana) la IA ha de prestarse específicamente a consideraciones morales. Pero ¿qué significa eso? Los expertos en ética disienten profundamente sobre el contenido de la misma, así como de su fundamento filosófico.

Cuanto más se piensa en las implicaciones de una “comunidad ética”, más problemática parece ser la noción de admitir las IA fuertes. Mucha gente intuye que la sugerencia misma es absurda.

MORALIDAD, LIBERTAD Y CONCIENCIA

Esa intuición surge en gran medida porque el concepto de responsabilidad moral se relaciona estrechamente con otros (voluntad consciente, libertad y conciencia) que contribuyen a nuestra noción de humanidad como tal.

La deliberación consciente hace que tengamos que rendir más cuenta moral sobre nuestras elecciones (aunque también se pueden criticar los actos irreflexivos). Al agente en cuestión, o a uno mismo, se le pueden atribuir elogios o culpa, y las acciones realizadas bajo graves coacciones son menos susceptibles de culpa que las realizadas libremente.

Estos conceptos son muy controvertidos hasta cuando se aplican a los seres humanos. Aplicarlos a máquinas parece inapropiado, sobre todo por las implicaciones que tendría para las interacciones entre seres humanos y ordenadores citadas en la sección anterior. Sin embargo, adoptar la perspectiva de que "la mente humana es una máquina virtual" puede ayudar a entender esos fenómenos en nuestro propio caso.

Los filósofos influidos por la IA (empezando por Marvin Minsky)[24] analizan la libertad desde el punto de vista de la complejidad cognitivo-motivacional. Observan que las personas son claramente "libres" de una forma en que los grillos, por ejemplo, no lo son. Las hembras de los grillos encuentran pareja mediante una respuesta refleja integrada (véase el capítulo V), pero una mujer que busca pareja tiene muchas estrategias a su disposición. También puede tener muchas otras motivaciones además de emparejarse que no pueden satisfacerse al mismo tiempo. Lo consigue, a pesar de todo, gracias a unos recursos computacionales (alias inteligencia) que los grillos no tienen.

Estos recursos, organizados por la conciencia funcional, incluyen el aprendizaje perceptual, la planificación anticipatoria, la asignación por defecto, la clasificación de prioridades, el razonamiento contrafáctico y la programación de acciones guiadas por la emoción. De hecho, Dennett usa estos conceptos (y muchísimos ejemplos ilustrativos) para explicar la libertad humana.[25] Así, gracias a la IA, entendemos mejor cómo es posible el libre albedrío.

El determinismo / indeterminismo es sobre todo una cortina de humo. Hay algún componente de indeterminismo en los actos humanos, pero no puede aparecer en el momento de la decisión porque menoscabaría la responsabilidad moral, aunque pueda afectar a las consideraciones que surgen durante la deliberación. El agente puede pensar o no en X o puede recordar Y, donde tanto X como Y incluyen hechos y valores morales. Por ejemplo, al que elige un regalo de cumpleaños le puede influir darse cuenta accidentalmente de algo que le recuerda que al receptor potencial le gusta el color morado o que defiende los derechos de los animales.

Todos estos recursos computacionales que se acaban de enumerar estarían al alcance de una IAF con capacidades humanas. Así que parece que nuestra IAF imaginaria tendría libertad, a menos que el libre albedrío deba incluir también conciencia fenoménica (y si rechazamos los análisis computacionales de eso). Si pudiéramos entender que la IAF tiene diversos motivos importantes para ella, entonces hasta podríamos hacer distinciones entre sus elecciones "libres" y sus elecciones "bajo coacción". No obstante, es un "si" condicional muy grande.

En cuanto a la identidad, los investigadores de IA destacan el papel de la computación recursiva, en la que un proceso puede operar sobre sí mismo. Esta conocida idea de la IA puede deshacer muchos dilemas filosóficos sobre el autoconocimiento (y el autoengaño).

Pero ¿de qué es conocimiento el "autoconocimiento"? Algunos filósofos niegan la realidad de la identidad, pero no los pensadores influidos por la IA, que la consideran un tipo específico de máquina virtual.

Para ellos, la identidad es una estructura computacional permanente que organiza y racionaliza las acciones del agente, especialmente las voluntarias que han sopesadas con detenimiento. (El autor de LIDA, por ejemplo, describe la identidad como "el contexto permanente de experiencia que organiza y estabiliza las experiencias a lo largo de muchos contextos locales diferentes"). La identidad no está presente en el bebé recién nacido, pero es una construcción para toda la vida (hasta cierto punto responsable del moldeado de uno mismo). Y su multidimensionalidad le permite variaciones considerables, lo que

genera voluntad individual reconocible y características propias personales.

Esto es posible porque la teoría de la mente del agente (que en principio interpreta el comportamiento de los demás) se aplica, reflexivamente, a los propios pensamientos y acciones. Les da sentido como motivos prioritarios, intenciones y objetivos. A su vez, estos se organizan por preferencias personales permanentes, relaciones personales y valores morales / políticos. Esta arquitectura computacional permite la construcción de la imagen de uno mismo (que representa el tipo de persona que uno cree ser) y de la imagen ideal de uno mismo (la clase de persona que uno querría ser) y de las acciones y emociones basadas en las diferencias entre ambas.

Dennett (muy influido por Minsky) llama a la identidad “centro de la gravedad narrativa”: una estructura (máquina virtual) que, al contar la historia de la vida de uno mismo, busca y genera una explicación de las propias acciones, en especial de las relaciones con los demás.[26] Esto, cómo no, deja margen para los autoengaños y la autoinvisibilidad de numerosos tipos.

De manera similar, Douglas Hofstadter describe las identidades como patrones abstractos autorreferenciales que surgen y regresan a la base sin significado de la actividad neuronal.[27] Estos patrones (máquinas virtuales) no son aspectos superficiales de la persona. Al contrario, para que exista la identidad solo se tiene que ejecutar ese patrón.

(Hofstadter añade que una persona muy amada puede seguir existiendo después de la muerte del cuerpo. La identidad de la persona “que se ha ido”, previamente ejecutada completamente en su cerebro, se ejecuta ahora a un nivel menos detallado en el cerebro de los supervivientes que lo amaban. Hofstadter insiste en que no se trata solo de “seguir viviendo” en la memoria de alguien o de que el sobreviviente adopte algunas características del otro como, por ejemplo, la pasión por la ópera. Es más que las dos identidades pre-muerte habrían interpretado las vidas mentales e ideas personales del otro tan profundamente que cada uno puede vivir literalmente en el otro. A través de su viudo, una madre muerta puede incluso experimentar de forma

consciente el crecimiento de sus hijos. Esta propuesta ilógica postula algo similar a la inmortalidad personal, aunque, cuando los supervivientes mismos mueren, la identidad perdida deja de ser ejecutada. Los filósofos “transhumanistas” vaticinan la inmortalidad personal duradera en los ordenadores: véase el capítulo VII).

En suma: decidir acreditarle a la IAF una inteligencia real equivalente a la humana (que incluya ética, libertad y conciencia) sería un gran paso con implicaciones prácticas significativas. Aquellos que rechazan la idea porque la consideran errónea en lo fundamental bien pueden estar en lo correcto. Por desgracia, ningún argumento filosófico que no sea controvertido puede respaldar esa intuición. No hay consenso en estas cuestiones, así que no hay respuestas fáciles.

MENTE Y VIDA

Todas las mentes que conocemos se encuentran en organismos vivos. Muchos, entre los que se encuentran los cibernéticos (véanse los capítulos I y V), creen que debe ser así; esto es, asumen que la mente presupone necesariamente la vida.

Los filósofos profesionales a veces hacen constar esto explícitamente, pero rara vez lo argumentan. Putnam, por ejemplo, dijo que es un “hecho indudable” que si un robot no está vivo entonces no puede ser consciente,[28] pero en vez de dar razones científicas, recurrió a “las reglas semánticas de nuestro lenguaje”. Ni siquiera la han demostrado fuera de toda duda los pocos que han defendido largo y tendido esta variable (como el filósofo medioambiental Hans Jonas[29] y recientemente el físico Karl Friston mediante su “principio de energía libre” cibernético).[30]

Asumamos, no obstante, que esta creencia común sea cierta. Si es así, entonces la IA puede llegar a tener inteligencia solo si también logra tener vida real. Debemos preguntarnos, entonces, si la “vida artificial fuerte” (vida en el ciberespacio) es posible.

No hay una definición de vida universalmente aceptada, pero por lo general se le atribuyen nueve rasgos característicos: autoorganización, autonomía, surgimiento, desarrollo, adaptación, capacidad de reacción,

reproducción, evolución y metabolismo. Los primeros ocho pueden entenderse en términos del procesamiento de la información, así que en principio en la IA / vida artificial se pueden encontrar ejemplos de todos ellos. La autoorganización, por ejemplo (que en un sentido amplio incluye a todos los demás) se ha logrado de varias formas (véanse los capítulos IV y V).

Pero el metabolismo es diferente.[31] Los ordenadores pueden replicarlo, pero no ejemplificarlo. Ni los robots autoensamblados ni la vida artificial virtual (en una pantalla) metabolizan de verdad. El metabolismo es el uso de sustancias bioquímicas e intercambios de energía para ensamblar y mantener el organismo, así que es irreduciblemente físico. Los defensores de la IA fuerte señalan que los ordenadores usan energía y que algunos robots tienen reservas de energía individuales que necesitan reabastecer de manera regular, pero muy lejos queda eso del uso flexible de ciclos bioquímicos entrelazados para construir el tejido corporal del organismo.

Así que, si el metabolismo es necesario para la vida, entonces la vida artificial fuerte es imposible. Y si la vida es necesaria para la mente, entonces la inteligencia artificial fuerte es imposible también. No importa lo impresionante que sea el desempeño de alguna IAF futura, no tendría inteligencia en realidad.

LA GRAN DIVISIÓN FILOSÓFICA

Los filósofos “analíticos” y los investigadores de IA también dan por sentado que es posible una psicología científica. De hecho, esa es la postura que se ha tomado a lo largo de este libro, incluido este capítulo.

Los fenomenólogos, no obstante, rechazan esa afirmación.[32] Argumentan que todos nuestros conceptos científicos surgen de la conciencia significativa, así que no se pueden emplear para explicarla. (El mismo Putnam acepta ahora esa postura).[33] Incluso sostienen que no tiene sentido postular la existencia de un mundo real independiente del pensamiento humano cuyas propiedades objetivas puedan ser desveladas por la ciencia.

Así que la falta de consenso sobre la naturaleza de la mente / inteligencia es aún más profunda de lo que he indicado hasta ahora.

No existe un argumento demoledor contra la visión de los fenomenólogos ni tampoco a su favor, ya que no existe un terreno común desde el que montar uno. Cada bando se defiende a sí mismo y critica al otro, pero usando argumentos cuyos términos de partida no han sido acordados mutuamente. La filosofía analítica y la fenomenológica dan interpretaciones diferentes sobre lo fundamental, incluso sobre conceptos básicos como razón y verdad. (El científico especializado en IA Brian Cantwell Smith ha propuesto una ambiciosa metafísica sobre computación, intencionalidad y objetos que tiene como objetivo respetar los conocimientos de ambas partes;[34] desafortunadamente, su fascinante argumento no es nada persuasivo).

Está discusión está sin resolver y quizá sea irresoluble. Para algunos, la postura de los fenomenólogos es “obviamente” correcta. Para otros, es “obviamente” absurda. Esa es otra razón más por la que nadie sabe seguro si una IAF podría ser de verdad inteligente.

VII

LA SINGULARIDAD

El futuro de la IA ha causado mucho revuelo desde su creación. Las predicciones demasiado entusiastas de (algunos) profesionales han emocionado y a veces aterrorizado a periodistas y comentaristas culturales. Ahora, el ejemplo más importante es la Singularidad: el momento que se propone como aquel en que las máquinas se vuelvan más inteligentes que los seres humanos.

Primero, se dice que la IA alcanzará un nivel de inteligencia igual al humano. (Se asume de manera tácita que sería inteligencia real: véase el capítulo VI). Poco después, la IA fuerte se transformará en IAS (“S” de sobrehumano), ya que los sistemas serán lo bastante inteligentes como para copiarse a sí mismos, y así sobrepasarnos en número, y mejorarse a sí mismos y así ser más inteligentes que nosotros. Los problemas y decisiones más importantes los abordarán los ordenadores.

Esta concepción es enormemente polémica. La gente discrepa sobre si podría pasar, si pasará, cuándo podría pasar y si sería algo bueno o malo.

Los que creen en la Singularidad (S-believers) argumentan que los avances en IA hacen que la Singularidad sea inevitable. Algunos la acogen con satisfacción. Pronostican que resolverá los problemas de la humanidad. La guerra, la enfermedad, el hambre, el aburrimiento y hasta la muerte personal… todos terminarán. Otros predicen el fin de la humanidad o, en todo caso, de la vida civilizada tal como la conocemos. Stephen Hawking (junto a Stuart Russell, coautor del manual más importante de la IA)[1] hizo mucho ruido a

nivel mundial en mayo de 2014 al decir que ignorar la amenaza de la IA sería “potencialmente nuestro peor error”.

Por el contrario, los escépticos de la Singularidad (S-skeptics) no esperan que suceda y desde luego no en el futuro inmediato. Admiten que la IA ofrece muchos motivos por los que preocuparse, pero no ven en ella una amenaza existencial.

LOS PROFETAS DE LA SINGULARIDAD

La idea de una transición de la IAF a la IAS se ha convertido en los últimos tiempos en un lugar común de los medios de comunicación, aunque se originó a mediados del siglo XX. Los iniciadores principales fueron “Jack” Good (un colega criptógrafo de Alan Turing en Bletchley Park), Vernor Vinge y Ray Kurzweil. (El propio Turing había previsto que “las máquinas tomaran el mando”, pero no dio detalles).

En 1965, Good predijo una máquina ultrainteligente que “superaría todas las actividades intelectuales de cualquier hombre por muy sabio que fuese”.[2] Como podría diseñar máquinas todavía mejores, “indudablemente [conduciría] a una explosión de inteligencia”. En aquella época, Good era moderadamente optimista: “La primera máquina ultrainteligente es el último invento que necesita hacer el hombre, ya que la máquina es lo bastante dócil como para decirnos cómo tenerla controlada”. Luego, sin embargo, afirmó que las máquinas ultrainteligentes nos destruirían.

Un cuarto de siglo más tarde, Vinge popularizó el término “Singularidad” (introducido en este contexto por John von Neumann en 1958).[3] Predijo “la singularidad tecnológica futura”, en la que todas las predicciones se descompondrán (como en el horizonte de sucesos de un agujero negro).

La Singularidad misma, concedió, puede pronosticarse, de hecho, es inevitable; pero entre las muchas consecuencias (desconocidas) podría estar la destrucción de la civilización e

incluso de la humanidad. Nos encaminamos a “descartar todas las normas previas, quizá en un abrir y cerrar de ojos, a una huida exponencial sin ninguna posibilidad de control”. Incluso aunque todos los gobiernos se dieran cuenta del peligro e intentasen evitarlo, no podrían, dijo Vinge.

El pesimismo de Vinge y (con el tiempo) el de Good fue contrarrestado por Kurzweil,[4] que aportó no solo un optimismo asombroso, sino también fechas.

Su libro, con el revelador título de La singularidad está cerca, sugiere que la IAF se logrará hacia 2030 y que hacia 2045 la IAS, (combinada con la nanotecnología y la biotecnología, habrá desterrado la guerra, la enfermedad, la pobreza y la muerte personal. También habrá provocado “una explosión de arte, ciencia y otras formas de conocimiento que […] le darán sentido a la vida”. A mediados de siglo, además, estaremos viviendo en realidades virtuales envolventes muchísimo más ricas y satisfactorias que el mundo real. Para Kurzweil, “la Singularidad” es singular de verdad y “cerca” significa cerca de verdad.

(Este hiperoptimismo a veces se atemperaba. Kurzweil enumera muchos riesgos existenciales, sobre todo de la IA asistida por la biotecnología. En cuanto a la IA misma, dice: “La inteligencia es intrínsecamente imposible de controlar […] Hoy es inviable concebir estrategias que aseguren completamente que la IA futura encarnará la ética y los valores humanos”).

El argumento de Kurzweil se fundamenta en “la ley de Moore”: la observación (de Gordon Moore, fundador de Intel) de que la potencia computacional disponible por un dólar se duplica cada año. (Las leyes de la física terminarán por darle alcance a la ley de Moore, pero no en un futuro inmediato). Como señala Kurzweil, cualquier crecimiento exponencial es notablemente contrario al sentido común y en este caso requiere que la IA avance a un ritmo inimaginable. Así que, como Vinge, insiste en que las expectativas basadas en experiencias anteriores no valen para casi nada.

PREDICCIONES COMPETITIVAS

A pesar de haber sido declarados casi inútiles, se suelen hacer pronósticos sobre la post-singularidad bastante a menudo. Hay multitud de ejemplos descabellados en la literatura, de los que aquí solo se pueden mencionar unos cuantos.

Los que creen en la Singularidad se dividen en dos campos: los pesimistas (como Vinge) y los optimistas (como Kurzweil). Por lo general están de acuerdo en que el paso de la IAF a la IAS sucederá antes de que termine el siglo, pero disienten en cómo de peligrosa será la IAS.

Por ejemplo, algunos vaticinan unos robots malvados que harán todo lo que esté a su alcance para desbaratar la vida y las esperanzas humanas (una figura común de la ciencia ficción y de las películas de Hollywood). La idea de que podríamos “desenchufarlos” si fuese necesario se niega específicamente. Las IAS, nos dicen, serían lo bastante astutas para convertir eso en imposible.

Otros arguyen que las IAS no tendrán intenciones malévolas, pero que de todas maneras serán peligrosísimas. No introduciríamos en ellas odio a los seres humanos y no hay ninguna razón por la que debieran desarrollarlo ellas. Más bien les seremos indiferentes, como nos son indiferentes a nosotros la mayoría de las especies no humanas. Pero su indiferencia, si nuestros intereses entran en conflicto con sus objetivos, podría ser nuestra ruina: el Homo sapiens terminaría como el dodo. En un experimento de pensamiento muy citado de Nick Bostrom, una IAS, en su resolución por hacer sujetapapeles, extraería de los cuerpos humanos los átomos para manufacturarlos.[5]

O también se puede pensar en la estrategia general sugerida a veces para protegerse contra las amenazas de la Singularidad: la contención. En este caso, se le impide a la IAS actuar directamente

sobre el mundo, aunque pueda percibir directamente el mundo. Se usa solo para responder preguntas (lo que Bostrom llama un "Oráculo"). No obstante, internet está incluido en el mundo y las IAS podrían provocar cambios indirectos aportando contenidos (hechos, falsedades, virus informáticos…).

Otra forma de pesimismo a propósito de la Singularidad predice que las máquinas nos pondrán a hacer el trabajo sucio, aunque vaya contra los intereses de la humanidad. Esta visión desdeña la idea de que podríamos "contener" los sistemas de IAS separándolos del mundo. Una máquina superinteligente, dicen, podría recurrir al chantaje o a la amenaza para convencer a alguno de los pocos seres humanos con los que se relacione para que haga cosas que ella no puede hacer directamente.

Esta inquietud en particular asume que la IAS habrá aprendido lo bastante sobre psicología humana como para saber qué chantajes o amenazas tienen probabilidad de funcionar, y quizá también qué individuos tienen probabilidad de ser más vulnerables a un cierto tipo de persuasión. A la objeción de que esta variable no es creíble, contestan que un vulgar soborno monetario o una amenaza de muerte funcionarían con casi cualquiera, así que la IAS no necesitaría una perspicacia psicológica a la altura de Henry James, ni necesitaría comprender en términos humanos lo que son en realidad la persuasión, el soborno y la amenaza. Bastaría con que supiera que es probable que introducir ciertos textos de PLN en un ser humano influya en su comportamiento de formas muy predecibles.

Algunas de las predicciones optimistas plantean retos aún mayores. Quizá las más llamativas sean las predicciones de Kurzweil de que viviremos en un mundo virtual y se eliminará la muerte personal. La muerte del cuerpo, si bien se retrasaría mucho (mediante biociencia asistida de IAS), continuaría sucediendo; pero el aguijón de la muerte podría extraerse descargando la personalidad y los recuerdos de cada persona en un ordenador.

Esta asunción filosóficamente problemática, que una persona podría existir ya fuese en silicio o en neuroproteína (véase el capítulo VI), se refleja en el subtítulo de este libro de 2005: Cuando los humanos trascendamos la biología. Kurzweil expresaba su visión “singulariana” (también llamada transhumanismo o poshumanismo) de un mundo que contuviera a personas que fueran en parte, o incluso por entero, no-biológicas.

Estos “ciborgs” transhumanistas, se dice, tendrán distintos implantes computerizados conectados directamente al cerebro y prótesis para las extremidades y/o los órganos de los sentidos. Se desterrarán la ceguera y la sordera, porque las señales visuales y auditivas se interpretarán mediante el sentido del tacto. No menos importante, la cognición racional (igual que los estados de ánimo) se potenciará con drogas de diseño específicas.

Las primeras versiones de tamañas tecnologías auxiliares ya están entre nosotros. Si proliferan como sugiere Kurzweil, nuestro concepto de humanidad cambiará profundamente. En vez de ser vistas como prótesis o añadidos útiles para el cuerpo humano, se verán como partes del cuerpo (trans)humanas. Las drogas psicotrópicas figurarán junto a sustancias naturales como la dopamina con relación al “cerebro”. Y la inteligencia, fuerza o belleza superiores de los individuos modificados genéticamente se verán como rasgos “naturales”. Se cuestionarán las opiniones políticas sobre el igualitarismo y la democracia. Hasta se podría desarrollar una nueva subespecie (¿o especie?) a partir de ancestros humanos lo suficientemente ricos como para explotar estas posibilidades.

En suma, se espera que la evolución tecnológica reemplace a la evolución biológica. Kurzweil ve la Singularidad como “la culminación de la fusión de nuestros pensamientos y existencia humanos con nuestra tecnología, que dará lugar a un mundo en el que no habrá distinción […] entre humano y máquina o entre realidad física y virtual”. (Se le perdonará al lector que necesite respirar hondo).

El transhumanismo es un ejemplo extremo de cómo la IA puede cambiar las ideas sobre la naturaleza humana. Una filosofía menos extrema que asimila la tecnología al concepto mismo de mente es “la mente extendida”, que considera que la mente se extiende por todo el mundo para abarcar los procesos cognitivos que dependen de él.[6] Aunque la noción de mente extendida ha sido muy influyente, el transhumanismo no. Algunos filósofos, críticos culturales y artistas[7] la han aprobado con entusiasmo. Sin embargo, no todos los que creen en la Singularidad la aceptan.

DEFENSA DEL ESCEPTICISMO

A mi modo de ver, los escépticos de la Singularidad tienen razón. La discusión del capítulo VI sobre si la mente es una máquina implicaba que no hay obstáculos en principio para que se dé la IA con nivel humano (posiblemente excepto la conciencia fenoménica). La cuestión es si es probable que se dé en la práctica.

Además de la implausibilidad intuitiva de muchas de las predicciones de la post-Singularidad y el casi disparate (en mi humilde opinión) de la filosofía transhumanista, los escépticos de la Singularidad tienen más argumentos de su parte.

La IA es menos prometedora de lo que mucha gente supone. En los capítulos II al V se han mencionado infinidad de cosas que la IA no puede hacer. Muchas requieren un sentido humano de la relevancia (y suponen de manera tácita la culminación de la web semántica: véase el capítulo II). Además, la IA se ha centrado en la racionalidad intelectual y ha ignorado la inteligencia social / emocional, y ni hablemos de la sabiduría. Una IAF capaz de interactuar plenamente con el mundo necesitaría también esas capacidades; añadamos la riqueza prodigiosa de la mente humana y la necesidad de buenas teorías psicológicas / computacionales sobre cómo funcionan y la perspectiva de una IA fuerte a nivel humano parece desvanecerse.

Aunque se pudiese realizar en la práctica, es dudoso que se obtenga la financiación necesaria. Los gobiernos ahora mismo están dedicando recursos tremendos a la transferencia mental (véase la siguiente sección), pero el dinero necesario para construir mentes humanas artificiales sería aún más.

Gracias a la ley de Moore, es indudable que se pueden esperar mayores avances de la IA, pero el aumento de potencia de los ordenadores y de la disponibilidad de datos (con el almacenamiento en la nube y sensores 24/7 por toda la Internet de las Cosas) no garantizan una IA de nivel humano. Malas noticias para los que creen en la Singularidad, porque la IAS requiere primero de la IAF.

Los que creen en la Singularidad ignoran las limitaciones de la IA actual. Simplemente, les da lo mismo, porque tienen un comodín: la noción de que los avances tecnológicos exponenciales están reescribiendo todos los libros de normas. Esto les da licencia para hacer predicciones a voluntad. A veces admiten que “a finales del siglo” las predicciones puede que no sean realistas. Sin embargo, insisten en que “nunca” es mucho tiempo.

Nunca es de hecho mucho tiempo, así que los escépticos de la Singularidad, incluyéndome a mí, puede que estén equivocados. No tienen ningún argumento arrollador, sobre todo si admiten la posibilidad de la IAF en principio (como es mi caso). Puede que hasta estén persuadidos de que la Singularidad, si bien muy demorada, terminará llegando.

Sin embargo, si sopesamos detalladamente la IA de última generación, hay buenas razones para respaldar las hipótesis (o las apuestas, si se prefiere) de los escépticos, más que las locas especulaciones de los que creen en la singularidad.

EMULACIÓN DEL CEREBRO COMPLETO

Los que creen en la Singularidad predicen un avance tecnológico exponencial en IA, biotecnología y nanotecnología y en la

cooperación entre ellas. De hecho, ya está pasando. Se están utilizando análisis de big data para avanzar en ingeniería genética, en desarrollo de medicamentos y en otros muchos proyectos de base científica (lo que vindicaba Ada Lovelace: véase el capítulo I). Del mismo modo, la IA y la neurociencia se están combinando para emular un cerebro completo (WBE, Whole Brain Emulation).

El objetivo de la WBE es mimetizar un cerebro real simulando sus componentes individuales (neuronas), junto con sus conexiones y su capacidad de procesar la información. Se espera que el conocimiento científico adquirido tenga muchas aplicaciones, incluyendo los tratamientos para patologías mentales, desde el alzheimer a la esquizofrenia.

Esta ingeniería inversa requerirá cálculos neuromórficos para modelar los procesos subcelulares como el paso de iones a través de la membrana celular (véase el capítulo IV).

El cálculo neuromórfico depende del conocimiento sobre la anatomía y fisiología de varios tipos de neuronas, pero la simulación cerebral también necesita evidencia detallada sobre conexiones neuronales específicas y su funcionalidad, incluida su coordinación. Buena parte de esto requerirá la mejora del escaneado cerebral mediante neuroprocesadores miniaturizados que observen a las neuronas individuales de manera continua.

Hay varios proyectos de WBE en curso, que sus patrocinadores suelen comparar con el Proyecto Genoma Humano o con la carrera espacial para llegar a la Luna. Por ejemplo, en 2013 la Unión Europea anunció el Proyecto Cerebro Humano, con un coste de mil millones de libras. Ese mismo año, el presidente Barack Obama anunció BRAIN, un proyecto de diez años financiado con tres mil millones de dólares del gobierno de Estados Unidos (más una significativa cantidad de dinero privado). Su objetivo primero es generar un mapa dinámico de la conectividad del cerebro del ratón y luego emular el cerebro humano.

Los gobiernos también financiaron otros intentos anteriores de emular una parte del cerebro. En 2005, Suiza patrocinó el proyecto Blue Brain, en principio para simular la columna cortical de una rata, pero con el objetivo a largo plazo de simular los millones de columnas del neocórtex humano. En 2008, la DARPA (Agencia de Proyectos de Investigación Avanzados de Defensa) destinó casi cuarenta millones de dólares a SyNAPSE (Systems of Neuromorphic and Plastic Scalable Electronics); en 2014 (y con cuarenta millones de dólares más) ya usaba chips con 5,4 mil millones de transistores, cada uno de ellos con un millón de unidades (neuronas) y 256 millones de sinapsis. Y Alemania y Japón colaboran en el uso de NEST (Tecnología de Simulación Neuronal) para desarrollar el ordenador K; en 2012, le llevaba cuarenta minutos simular un segundo del 1% de la actividad cerebral real, en la que participan 173 mil millones de "neuronas" y 10,4 billones de "sinapsis".

Como es tan caro, el WBE mamífero es escaso, pero en todo el mundo se están dando innumerables intentos de cartografiar cerebros mucho más pequeños (en mi universidad, se centran en el cerebro de las abejas), que pueden proporcionar a los neurocientíficos observaciones que ayuden al WBE a escala humana.

Dados los progresos de hardware que ya se han alcanzado (por ejemplo, los chips de SyNAPSE), más la ley de Moore, la predicción de Kurzweil de que hacia 2020 existirán ordenadores que igualen en crudo la capacidad de procesamiento del cerebro humano es plausible, pero su convencimiento de que serán equiparables a la inteligencia humana alrededor de 2030 ya es otra cuestión.

Lo que es crucial en este caso es la máquina virtual (véanse los capítulos I y VI). Algunas máquinas virtuales se pueden implementar solo en un hardware tremendamente potente, así que quizá se necesiten chips mega-transistorizados. Pero ¿qué cálculos llevarán a cabo? Dicho de otro modo, ¿qué máquinas virtuales se implementarán en ellos? Para equiparar a la inteligencia humana (o

incluso a la inteligencia de un ratón), tendrán que ser informativamente impactantes de maneras que los psicólogos computacionales todavía no terminan de entender.

Supongamos (lo que creo que es poco probable) que cada neurona del cerebro humano sea, en algún momento, cartografiada. Esto en sí no nos dirá lo que está haciendo (el nematodo C. elegans, un gusano minúsculo, tiene solo 302 neuronas, cuyas conexiones se conocen con toda precisión, pero no podemos identificar siquiera si una sinapsis sirve para excitarlo o para inhibirlo).

Ya tenemos un mapa muy detallado del córtex visual entre neuroanatomía y función psicológica, pero no del neocórtex en general. Concretamente, no sabemos mucho de lo que hace el córtex frontal, esto es, qué máquinas virtuales se implementan en él. Esta cuestión no es relevante en la WBE a gran escala. El Proyecto Cerebro Humano, por ejemplo, ha adoptado un planteamiento decididamente ascendente: observar la anatomía y la bioquímica y copiarlas. Las cuestiones que van de arriba hacia abajo en la escala jerárquica sobre funciones fisiológicas que puede que se sustenten en el cerebro quedan al margen (muy pocos neurocientíficos cognitivos se dedican a ellas). Incluso si se consiguiera un modelo anatómico completo y se observaran con todo detalle los mensajes químicos, no se podría responder a esas cuestiones que van de lo general a lo concreto.

Para responder haría falta una gran variedad de conceptos computacionales. Además, una cuestión clave es la arquitectura computacional de la mente (o mente-cerebro) en su conjunto. Hemos visto en el capítulo III que la planificación de las acciones de criaturas con motivaciones múltiples requiere la programación de mecanismos complejos, como los que proporcionan las emociones. Y la discusión sobre LIDA en el capítulo VI dio una idea de la enorme complejidad de los procesos corticales. Hasta una actividad cotidiana como comer con tenedor y cuchillo requiere que se integren muchas máquinas virtuales, algunas que se encarguen de

los objetos físicos (músculos, dedos, utensilios, varios tipos de sensores), otras de las intenciones, planes, expectativas, deseos, convenciones sociales y preferencias. Para entender cómo es posible esa actividad, necesitamos, además de datos neurocientíficos sobre el cerebro, teorías computacionales detalladas sobre los procesos psicológicos involucrados.

En suma, considerada como vía para comprender la inteligencia humana, es probable que la WBE ascendente fracase. Nos puede enseñar mucho sobre el cerebro y puede ayudar a que los científicos de la IA desarrollen más aplicaciones prácticas, pero la idea de que a mitad de siglo la WBE habrá explicado la inteligencia humana es ilusoria.

LO QUE DEBERÍA PREOCUPARNOS

Si los escépticos tienen razón y no va a haber Singularidad, de ahí no se desprende que no haya nada por lo que preocuparse. La IA ya plantea motivos de preocupación. Con los progresos futuros seguramente surgirán más, así que la ansiedad sobre los peligros de la IA a largo plazo no está totalmente fuera de lugar. En concreto, hay que prestar atención también a sus efectos a corto plazo.

Algunas preocupaciones son muy generales. Por ejemplo, cualquier tecnología se puede utilizar para el bien o para el mal. Los malvados usarán cualquier herramienta disponible (y a veces financiarán la elaboración de medios nuevos) para hacer cosas malvadas. (CYC, por ejemplo, podría serles útil a los malhechores: sus desarrolladores ya están pensando en cómo limitar el acceso a todo el sistema cuando salga al mercado; véase el capítulo II). Así que debemos tener mucho cuidado con lo que inventamos.

Como señala Stuart Russell, esto implica algo más que ser prudentes con nuestros objetivos. Si hay diez parámetros relevantes para un problema y optimizar estadísticamente el aprendizaje automático (véase el capítulo II) tiene en cuenta solo seis, entonces los otros cuatro podrán ser llevados al extremo y seguramente así

será. Por tanto, es necesario que estemos atentos a qué tipo de datos se están usando.

Esta preocupación general concierne al problema del marco (véase el capítulo II). Como el pescador del cuento, a quien le fue concedido el deseo de que su hijo soldado volviese a casa y se hizo realidad cuando se lo trajeron en un ataúd, podríamos llevarnos una sorpresa desagradable si los sistemas potentes de IA no tuviesen nuestro mismo sentido de la relevancia.

Por ejemplo, cuando un sistema de detección temprana durante la Guerra Fría recomendó un ataque defensivo sobre la URSS, el desastre se advirtió solo gracias al sentido de relevancia, tanto político como humanitario, de sus operadores,[8] que juzgaron que los soviéticos no habían sido especialmente escandalosos en la ONU en los últimos tiempos y temían las horrendas consecuencias de un ataque nuclear. Así que, violando el protocolo, ignoraron el aviso automático. Han ocurrido muchos otros cuasi incidentes nucleares; algunos, hace poco. Por lo general, lo único que impidió la escalada fue el sentido común de los seres humanos.

Por otra parte, siempre es posible que ocurra un error humano. A veces, es comprensible (el accidente de Three Mile Island fue peor porque los seres humanos hicieron caso omiso del ordenador, pero las condiciones físicas a las que se enfrentaban eran muy insólitas), pero puede ser terriblemente inesperado. La alerta durante la Guerra Fría del párrafo anterior sucedió porque alguien se había olvidado de programar los años bisiestos en el calendario, así que la luna estaba en el sitio “equivocado”. Con más razón, entonces, hay que someter a pruebas (en la medida de lo posible) y demostrar la fiabilidad de los programas de IA antes de utilizarlos.

Otras inquietudes son más específicas y algunas deberían estar perturbándonos ya.

Una amenaza primordial es el desempleo tecnológico. Muchos trabajos manuales y de nivel administrativo bajo han desaparecido.

Otros les seguirán (aunque los trabajos manuales que requieren destreza y adaptación no desaparecerán). La mayoría de la carga, acarreo y transporte de los almacenes la pueden hacer los robots y los vehículos sin conductor eliminarán puestos de trabajo.

Los puestos medios de gestión administrativa también peligran. Muchos profesionales ya utilizan sistemas de IA como asistentes. No falta mucho para que los trabajos (de derecho y contabilidad, por ejemplo) que requieren mucho tiempo de investigación de regulaciones y precedentes pueden ser en gran parte asumidas por la IA. Algunas tareas más exigentes, incluidas muchas de medicina y ciencia, no tardarán en verse afectadas. Los trabajos, aunque no se pierdan, necesitarán menos cualificación y la formación profesional se resentirá: ¿cómo aprenderán los jóvenes a juzgar con sensatez?

Aunque algunos trabajos en el campo legal serán redundantes, los abogados saldrán ganando con la IA, porque en ella acechan multitud de trampas legales. Si algo saliera mal, ¿quién será el responsable: el programador, el mayorista, el minorista o el usuario? ¿Y se podría demandar alguna vez a un profesional humano por no usar un sistema de IA? Si se ha demostrado (ya sea matemática o empíricamente) que el sistema es altamente fiable, un litigio de ese calibre sería muy probable.

Aparecerán sin duda nuevos tipos de trabajo, pero es dudoso que vayan a ser equivalentes en cuanto a cantidad, formación asequible y/o poder adquisitivo (como sucedió tras la revolución industrial).[9] Se avecinan serios desafíos sociopolíticos.

Los cargos públicos están menos amenazados, pero también se verán comprometidos. En un mundo ideal, la oportunidad para multiplicar y mejorar las actividades que ahora están infravaloradas sería acogida con entusiasmo, pero esto no está garantizado.

Por ejemplo, la educación está empezando a incluir asistentes de IA personales y/o por internet (como los COMA –curso online

masivo abierto– o MOOC por sus siglas en inglés, que ofrecen clases magistrales de estrellas académicas) que bajan el nivel del trabajo de muchos profesores humanos. Ya se dispone de psicoterapeutas computerizados, a un coste mucho menor que los terapeutas humanos (algunos son sorprendentemente útiles, por ejemplo, para reconocer la depresión). No obstante, no tienen regulación alguna. Y vimos en el capítulo III que el cambio demográfico estimula la investigación en el campo potencialmente lucrativo de los “cuidadores” artificiales para ancianos y de “niñeras robot”.[10]

Al margen de los efectos que tendrá en el desempleo, el uso de sistemas de IA sin empatía en contextos tan esencialmente humanos es tan arriesgado en la práctica como éticamente dudoso. Muchos “acompañantes computerizados” están diseñados para el uso de ancianos y/o personas con discapacidad que tienen un contacto mínimo con los pocos seres humanos con los que se encuentran. Están destinados a ser fuente no solo de ayuda y entretenimiento sino también de conversación, convivencia y consuelo emocional. Incluso si a la persona vulnerable le hiciese más feliz dicha tecnología (como a los usuarios de Paro), se estaría traicionando su dignidad humana de manera insidiosa. (Las diferencias culturales son importantes en este caso: la actitud hacia los robots difiere ampliamente en Japón y en occidente, por ejemplo).

A los ancianos les puede gustar hablar de sus recuerdos con un acompañante artificial, pero ¿sería eso una conversación de verdad? Podría ser un recordatorio gratificante que desencadenase reconfortantes momentos de nostalgia. No obstante, se puede proporcionar ese beneficio al usuario sin inducirle la ilusión de la empatía. Muchas veces, incluso en terapias emocionalmente tensas, lo que la persona desea por encima de todo es que se le reconozca su valentía y/o su sufrimiento, pero que surja de un entendimiento mutuo de la condición humana. Estamos estafando a la persona al ofrecerle solo un simulacro superficial de comprensión.

Incluso si el usuario padece de demencia moderada, su “expectativa” del agente de IA probablemente será mucho más rica que la que el agente tenga del modelo humano. ¿Qué pasaría entonces si el agente no reacciona según lo que se espera, o que se necesita, cuando la persona recuerda alguna pérdida personal dolorosa (de un hijo quizá)? Las expresiones convencionales de comprensión del acompañante no ayudarían y podrían causar más mal que bien. Mientras, se habría provocado que la persona se afligiera sin poder obtener consuelo inmediato.

Otra inquietud es si el acompañante debería callarse algunas veces o contar una mentira piadosa. Que dijese verdades implacables (y los silencios súbitos) podría molestar al usuario, pero la diplomacia requeriría procesamiento de lenguajes naturales (PLN) muy avanzado, más un modelo sutil de psicología humana.

En cuanto a los robots niñeras (y si ignoramos los problemas de seguridad), el uso excesivo de sistemas de IA por parte de bebés y niños pequeños podría alterar su desarrollo social y/o lingüístico.

Las parejas sexuales artificiales, además de protagonizar películas (en Her, por ejemplo), ya se están comercializando. Algunas tienen reconocimiento de habla y pueden hablar y moverse seductoramente. Son una influencia más, y aumentada, de las que ya nos brinda internet para vulgarizar la experiencia sexual humana (y refuerzan la cosificación sexual de la mujer). Muchos opinadores (incluyendo a algunos científicos especializados en IA) han escrito sobre sus relaciones sexuales con robots en unos términos que revelan un concepto del amor íntimo sumamente superficial que se presta a la confusión con la lujuria, la obsesión sexual y una mera familiaridad cómoda.[11] No obstante, es poco probable que estas advertencias sean efectivas. Dada la enorme rentabilidad de la pornografía en general, hay pocas esperanzas de impedir los futuros “adelantos” en muñecas sexuales con IA.

La privacidad es otro asunto espinoso, que cada vez se vuelve más polémico, porque se deja que motores de búsqueda y

aprendizaje de IA anden sueltos por los datos recopilados de aparatos multimedia personales y los sensores domésticos o portátiles. (Google ha patentado hace poco un oso de peluche robótico, con cámaras en los ojos, micrófonos en las orejas y altavoces en la boca. Podrá comunicarse con los padres y con el niño y, se quiera o no, también con recopiladores de datos invisibles).

Hace tiempo que la ciberseguridad es un problema. Cuanto más se mete la IA en nuestro mundo (la mayoría de las veces de forma nada transparente), más grave será. Una defensa contra la toma de poder de la IAS sería encontrar formas de escribir algoritmos que no pudiesen ser pirateados ni alterados (un objetivo de la “IA amigable”: véase la siguiente sección).

Las aplicaciones militares también suscitan inquietudes. Los robots dragaminas son bienvenidos, pero ¿y los soldados robóticos o las armas robóticas? Los drones actuales son inducidos por seres humanos, pero incluso así podrían aumentar los sufrimientos al ampliar la distancia humana (no solo geográfica) entre el operario y su blanco. Debemos esperar que a los drones futuros no les esté permitido decidir quién o qué debería ser el blanco. Incluso confiar en que reconocerán un blanco (elegido por seres humanos) genera inquietudes éticamente perturbadoras.

LO QUE SE ESTÁ HACIENDO AL RESPECTO

Ninguna de estas inquietudes es nueva, aunque pocos de los que trabajan en IA les han prestado mucha atención hasta ahora.

Varios pioneros de la IA estudiaron sus implicaciones sociales en un encuentro en el lago Como en 1972, pero John McCarthy se negó a unirse a ellos y dijo que era demasiado pronto para especular. Pocos años después, el científico informático Joseph Weizenbaum publicó un libro con el subtítulo Del juicio a la computación en el que lamentaba la “obscenidad” de confundir los dos, pero la comunidad de la IA lo rechazó con desdén.[12]

Hubo algunas excepciones, por supuesto. Por ejemplo, el primer libro que presentó una visión de conjunto de la IA incluía un capítulo final sobre “Relevancia social”[13] y el CPSR (Profesionales Informáticos pro-Responsabilidad Social) se fundó en 1983 (gracias a los esfuerzos de, entre otros, el creador de SHRDLU Terry Winograd: véase el capítulo III). Se hizo sobre todo para advertir de la falta de seriedad de la tecnología de Star Wars; el científico informático David Parnas incluso se dirigió al senado de Estados Unidos a este respecto. Conforme fueron disminuyendo las preocupaciones de la Guerra Fría, la mayoría de los profesionales de la IA se involucraron menos en su campo. Solo unos pocos, como Noel Sharkey de la universidad de Sheffield (un experto en robótica que preside el Comité Internacional de Control de Armas Robóticas),[14] más algunos filósofos de la IA, por ejemplo, Wendell Wallach de Yale[15] y Blay Whitby de Sussex[16] siguieron enfocándose en asuntos sociales o éticos durante años.

Ahora, debido tanto a la práctica como a la promesa de la IA, los recelos se han vuelto más apremiantes. Dentro del sector (y, hasta cierto punto, fuera de él) se le está prestando más atención a las implicaciones sociales.

Algunas reacciones importantes no tienen nada que ver con la Singularidad. Por ejemplo, la ONU y el Observatorio de Derechos Humanos llevan mucho tiempo promoviendo un tratado (aún no firmado) que prohíba las armas totalmente autónomas, como los drones que seleccionan su objetivo. Y algunos organismos profesionales consolidados han revisado las prioridades de sus investigaciones y/o sus códigos de conducta, pero hablar de la Singularidad ha atraído al debate a colaboradores adicionales.

Muchos (los que creen en la Singularidad y los escépticos por igual) sostienen que, aunque la probabilidad de la Singularidad sea extremadamente pequeña, las posibles consecuencias son tan graves que deberíamos empezar a tomar precauciones ya. A pesar de la afirmación de Vinge de que no se puede hacer nada a

propósito de esta amenaza existencial, se han creado varias instituciones para protegerse de ella.

Entre ellas está el Centro para el Estudio del Riesgo Existencial (CSER) de Cambridge, el Instituto para el Futuro de la Humanidad (FHI) de Oxford, el Instituto de la Vida Futura (FLI) de Boston y el Instituto de Investigación de la Inteligencia de las Máquinas (MIRI) de Berkeley. Estas organizaciones están en su mayor parte financiadas por filántropos del ámbito de la IA. Por ejemplo, Jaan Tallinn, uno de los desarrolladores de Skype, fue el cofundador de CSER y de FLI. Ambas instituciones, además de comunicarse con los profesionales de la IA, intentan alertar de los peligros a los responsables políticos y a otros miembros influyentes de la sociedad.

El presidente de la American Association for IA (Eric Horwitz) organizó un pequeño comité en 2009 para hablar de las precauciones que podrían ser necesarias para guiar o incluso retrasar el trabajo socialmente problemático de la IA. La reunión tuvo lugar, con toda intención, en Asilomar, California, donde los genetistas habían acordado unos años antes una moratoria sobre ciertas investigaciones genéticas. Sin embargo, como miembro del grupo, tuve la impresión de que no a todos les preocupaba seriamente el futuro de la IA. El informe ulterior no tuvo gran repercusión en los medios.

Con el mismo motivo, pero con más concurrencia, se convocó el encuentro del Instituto de la Vida Futura (FLI) y el Centro para el Estudio del Riesgo Existencial (CSER) (con arreglo a las normas de Chatham House y sin periodistas presentes) en Puerto Rico en enero de 2015. El organizador, Max Tegmark, era uno de los que habían firmado la carta amenazadora junto a Russell y Hawking seis meses antes. No fue una sorpresa, pues, que el ambiente fuese considerablemente más apremiante que en Asilomar. Este encuentro dio lugar de inmediato a nuevos y generosos fondos (del millonario de internet Elon Musk) para investigar sobre la seguridad de la IA y la IA ética, más una carta abierta de advertencia firmada

por miles de profesionales de la IA que tuvo una amplia difusión en los medios.

Poco después, una segunda carta abierta redactada por Tom Mitchell y varios investigadores destacados más advertía contra la creación de armas autónomas que elegirían y atacarían sus objetivos sin intervención humana. Los signatarios esperaban "impedir que comenzara una carrera armamentística global de IA". La carta se presentó en la Conferencia Internacional de IA de julio de 2015, con la firma de casi 3.000 científicos y 17.000 personas de campos afines y tuvo gran repercusión mediática.

La reunión de Puerto Rico también produjo una carta abierta (en junio de 2015) de los economistas del MIT Erik Brynjolfsson y Andy McAfee, esta vez dirigida a legisladores, emprendedores, empresarios y economistas profesionales. En ella advertían sobre las potenciales implicaciones económicas drásticas de la IA, y hacían algunas sugerencias sobre legislación pública que podrían mejorar (aunque no anular) los riesgos.

Estos esfuerzos de la comunidad de la IA están convenciendo a las fuentes de financiación gubernamentales de Estados Unidos de la importancia de los temas sociales y éticos. El departamento de Defensa y la Fundación Nacional para la Ciencia de Estados Unidos han dicho hace poco que están dispuestos a financiar las investigaciones,[17] pero este apoyo no es del todo nuevo: el interés "gubernamental" lleva aumentando unos cuantos años.

Por ejemplo, dos Consejos de Investigación del Reino Unido patrocinaron en 2010 una "residencia robótica" interdisciplinaria, en parte para esbozar un código de conducta destinado a los especialistas en robótica. Se acordaron cinco "principios", dos de los cuales trataban de inquietudes que ya hemos analizado: "(1) Los robots no deben ser diseñados como armas, excepto por razones de seguridad nacional" y "(4) Los robots son artefactos manufacturados: la ilusión de respuesta emocional e intención no debería utilizarse para aprovecharse de usuarios vulnerables".

Dos más pusieron la responsabilidad moral directamente sobre hombros humanos: “(2) Los seres humanos, no los robots, son los agentes responsables…” y “(5) Debería ser posible averiguar quién es [legalmente] responsable de cualquier robot”. El grupo se abstuvo de intentar actualizar las “tres leyes de la robótica” de Isaac Asimov (Un robot no hará daño a un ser humano, y debe obedecer las órdenes humanas y proteger su propia existencia siempre que las dos últimas leyes no entren en contradicción con la primera). Insistieron en que las “leyes” que haya las tienen que seguir los diseñadores / constructores humanos, no el robot.

En mayo de 2014, una iniciativa académica financiada por la Marina de Estados Unidos (7,5 millones de dólares para cinco años) fue aclamada en todos los medios. Es un proyecto de cinco universidades (Yale, Brown, Tufts, Georgetown y el Instituto Rensselaer) que tiene el objetivo de desarrollar la “competencia ética” para robots. En él están implicados psicólogos cognitivos y sociales y filósofos éticos, así como programadores e ingenieros de IA.

La intención de este grupo interdisciplinario no es proporcionar una lista de algoritmos éticos (comparable a las leyes de Asimov) ni dar prioridad a una metaética particular (por ejemplo, el utilitarismo), ni tampoco definir un conjunto de valores morales no competidores. Lo que pretende es desarrollar un sistema computacional capaz de razonamiento ético (y debate ético) en el mundo real, ya que los robots autónomos no solo seguirán instrucciones (y menos reaccionando inflexiblemente a pies “situados”: véase el capítulo V), sino que tomarán decisiones deliberativas. Si un robot participa en una búsqueda de rescate, por ejemplo, ¿a quién debería desalojar / rescatar primero? O si proporciona compañía social, ¿cuándo –si se da el caso– debería no contarle la verdad al usuario?

El sistema propuesto incorporaría percepción, acción motora, procesamiento de lenguajes naturales (PLN), razonamiento (tanto

deductivo como analógico) y deducción. Esta última incluiría pensamiento emocional (que puede anunciar acontecimientos importantes, así como programar objetivos conflictivos: véase el capítulo III); dispositivos robóticos de “protesta y malestar”, que podrían influir en las decisiones que tomen las personas que interactúen con ellos y el reconocimiento de las emociones de los seres humanos de su entorno. Y, tal como declara el anuncio oficial, el robot podría incluso “exceder” la competencia ética normal (esto es, la humana).

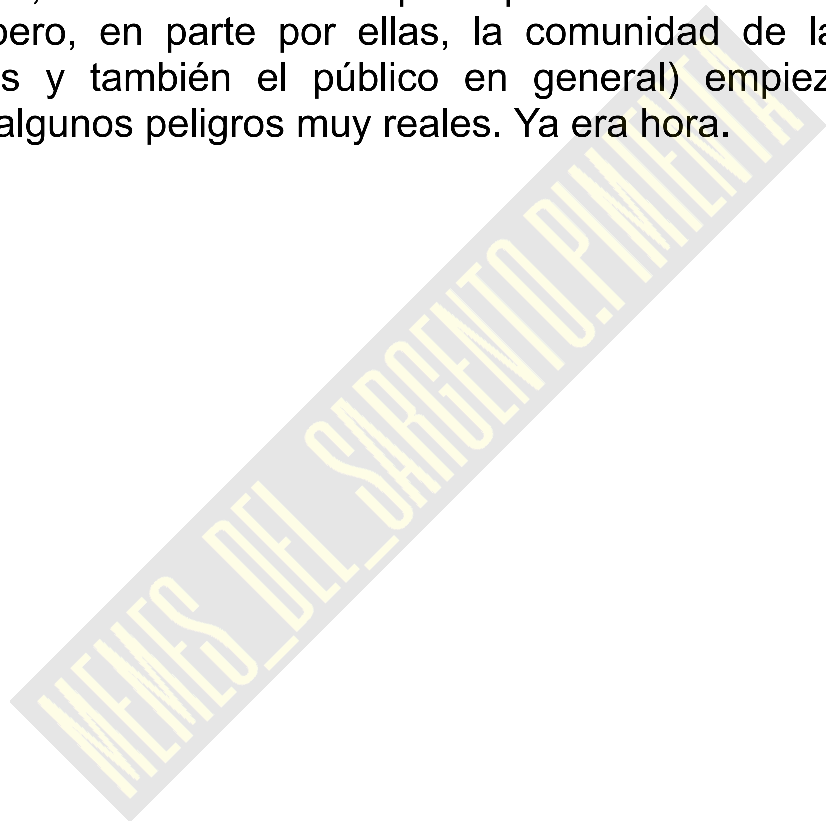
Dados los obstáculos comentados en los capítulos II y III, más las dificultades relativas a la ética específicamente (véase el capítulo VI), para conseguir la IAF bien podría poner en duda que esta tarea sea factible. Sin embargo, el proyecto podría valer la pena, ya que estudiar los problemas del mundo real (como los dos ejemplos tan diferentes de los que se ha hablado) nos puede advertir de los muchos riesgos que tiene utilizar inteligencia artificial en situaciones moralmente problemáticas.

Además de este esfuerzo institucional, son cada vez más los científicos se proponen como objetivo lo que Eliezer Yudkowsky llama “IA amigable”,[18] una IA con efectos positivos para la humanidad, tan seguros como útiles. En ella se deberían incluir algoritmos inteligibles, confiables, robustos y que fallen con elegancia, si acaso fallan. Deberían ser transparentes, predecibles y no vulnerables a la manipulación de los hackers, y si se puede demostrar su fiabilidad mediante lógica o las matemáticas, en vez de con pruebas empíricas, mucho mejor.

Los seis millones de dólares donados por Musk en la reunión de Puerto Rico dieron lugar de inmediato a una “convocatoria de propuestas” sin precedentes del Future of Life Institute (seis meses después, se habían financiado treinta y siete proyectos). Esta convocatoria, dirigida a expertos en “políticas estatales, leyes, ética, economía o educación y compromiso con la sociedad” así como de IA, solicitaba “proyectos de investigación dirigidos a maximizar los beneficios sociales futuros de la inteligencia artificial al mismo

tiempo que se evitan sus riesgos potenciales" y "limitados a la investigación explícitamente orientada no al objetivo habitual de mejorar la capacidad de la IA, sino de hacer que sea más robusta y/o beneficiosa…". Ese llamamiento a la IA amigable podría haber tenido lugar de todas formas, quizá, pero la huella de la Singularidad era visible: "Se dará prioridad a las investigaciones cuyo objetivo sea que la IA siga siendo robusta y beneficiosa, aunque llegue a superar considerablemente las capacidades actuales…", decía.

En suma, las visiones casi apocalípticas sobre la IA futura son ilusorias, pero, en parte por ellas, la comunidad de la IA (y los legisladores y también el público en general) empieza a darse cuenta de algunos peligros muy reales. Ya era hora.

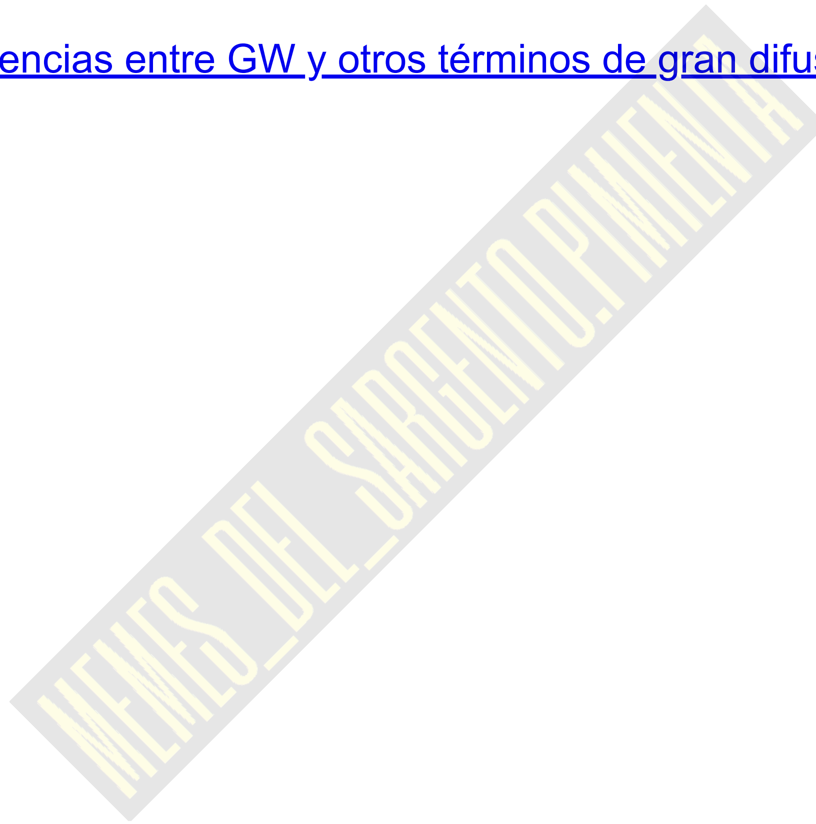


AGRADECIMIENTOS

Agradezco a los siguientes amigos sus útiles consejos (los errores son, por supuesto, míos): Phil Husbands, Jeremy Reffin, Anil Seth, Aaron Sloman y Blay Whitby. Y le agradezco a Latha Menon su comprensión y paciencia.



LISTA DE ILUSTRACIONES



NOTAS

I. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

1 Lovelace (1843).

2 Turing (1936).

3 Turing (1950).

4 McCulloch y Pitts (1943).

5 Samuel (1959).

6 Newell y Simon (1956).

7 Newell, Shaw y Simon (1959).

8 Uttley (1956, 1959).

9 Beurle (1956).

10 Pitts y McCulloch (1947).

11 Wiener (1948).

12 Turing (1952).

13 Se habla de todos estos escritores en Boden (2006), 4, pp. v-viii.

14 Craik (1943).

15 von Neumann (1951).

16 Ashby (1947, 1952).

17 Grey Walter (1950).

18 Selfridge (1959).

19 Blake y Uttley (1959).

20 Rosenblatt (1958, 1962).

21 Minsky y Papert (1969).

II. LA INTELIGENCIA GENERAL ES EL SANTO GRIAL

1 McCarthy (1959, 1963).

2 Yudkowsky (2008).

3 Gigerenzer (2004).

4 Hsu (2002).

5 Boukhtouta et al. (2005).

6 Boukhtouta et al. (2005).

7 Sahota y Mackworth (1994).

8 Michalski y Chilausky (1980).

9 Bengio et al. (2003); Collobert et al. (2011); Mikolov et al. (2013); Bahdanau et al. (2015).

10 Schank y Abelson (1977).

11 McCarthy (1980, 1986, p. 198).

12 McCarthy y Hayes (1969).

13 Marr (1982); Gibson (1979).

14 Vincze et al. (2014).

15 Véase Boden (2006), 14 p. ix, e.

16 Véase, por ejemplo: Garcia y Delakis (2004); Krizhevsky et al. (2012); Taigman et al. (2014); Xu et al. (2015).

17 Sloman (1971, 1975).

18 McCarthy y Hayes (1969).

19 Shanahan (1997).

20 Véase Boden (2006), 13, pp. iii, d-e.

21 Hutchins (1995).

22 Mitchell (1997, 2006).

23 En especial, Geoffrey Hinton y el equipo DeepMind dirigido por Demis Hassabis (que desarrolló el jugador de Atari descrito en el libro y el sistema que venció al campeón europeo de Go en 2016).

24 Para un informe sobre los avances recientes en aprendizaje profundo, véase LeCun, Bengio y Hinton (2015).

25 Mnih y Hassabis et al. (2015).

26 Véase LeCun, Bengio y Hinton (2015, pp. 442).

27 Newell y Simon (1972).

28 Laird et al. (1987); Newell (1990).

29 Anderson (1983); Anderson y Lebiere (2003).

30 Lenat (1995); Deaton y Lenat et al. (2005); véase http://www.cyc.com.

31 Sun et al. (2005).

32 Minsky (1985, 2006); Sloman (sin fecha).

33 Minsky (1956/1961).

III. LENGUAJE, CREATIVIDAD, EMOCIÓN

1 Davey (1978).

2 Hinton, Deng, Yu et al. (2012); Graves, Mohamed y Hinton (2013).

3 Winograd (1972).

4 Véase Hutchins (1985, pp. 164-7).

5 Taube (1961).

6 Wilks (2005, p. 266).

7 Bengio et al. (2003); Collobert et al. (2011); Mikolov et al. (2013).; Bahdanau et al. (2015).

8 Collobert (2011).

9 Bartlett et al. (2014).

10 Baker (2012).

11 Boden (1990/2004).

12 Kolodner (1992).

13 El modelo analógico del procesador en paralelo CopyCat está menos constreñido, pero también está restringido (Hofstadter y Mitchell 1997).

14 Cope (2005).

15 Koning y Eizenberg (1981).

16 Boden y Edmonds (2009).

17 Colton (2012).

18 Cohen (1995, 2002).

19 Todd y Latham (1992); McCormack (2004). Véase también Whitelaw (2004).

20 Simon (1967).

21 Colby (1963); Colby et al. (1971). De la obra de Colby se habla ampliamente en Boden (1977), pp. 21-63 y 97-106.

22 Wilks (2010).

23 Picard (1997, 1999); Sloman (1999).

24 Minsky (2006); Sloman (sin fecha).

IV. REDES NEURONALES ARTIFICIALES

1 Rosenblatt (1958, 1962); véase el capítulo I.

2 Véase Boden (2006), capítulo 14.

3 Clark (1989, 1993); Churchland (1989).

4 Rumelhart y McClelland (1986b); Pinker y Prince (1988); Sampson (2005).

5 Clark y Karmiloff-Smith (1993); Clark y Thornton (1997); Thornton (2000).

6 Rumelhart y McClelland (1986a); McClelland y Rumelhart (1986).

7 Clark (1989, 1993).

8 Dayan y Abbott (2001): capítulo 8.

9 Hinton y Sejnowski (1985).

10 Véase la revisión histórica de LeCun, Bengio y Hinton (2015).

11 Sutskever, Martens y Hinton (2011); Sutskever, Vintals y Le (2014).

12 Hinton et al. (2006).

13 Minsky y Papert (1969).

14 Se describe este trabajo en Boden (2006), 12.v.

15 Minsky y Papert (1988).

16 Se describen otros en Boden (2006), 11, pp. ii-iv, incluidos los escándalos relacionados con Joseph Weizenbaum, Herbert Dreyfus, James Lighthill y la “estupidez natural” de Drew McDermott.

17 Neisser (1963).

18 Haugeland (1978).

19 Philippides et al. (1998, 2005).

20 Mackay (1949/1959).

21 Minsky (1956/1961).

22 Shallice y Cooper (2011).

23 Cooper et al. (1996, 2005).

24 Shallice and Cooper (2011).

25 Minsky (1985).

V. LOS ROBOTS Y LA VIDA ARTIFICIAL

1 Véase Boden (1996).

2 Beer (1990).

3 Webb (1996).

4 Hutchins (1995).

5 Brooks (1991).

6 Kirsh (1991).

7 Sahota y Mackworth (1994).

8 Cliff et al. (1993).

9 Harvey et al. (1994).

10 Bird y Layzell (2002).

11 Ray (1991)

12 Turing (1952).

13 Turk (1991).

14 Goodwin (1994); Langton (1991).

15 Gardner (1970); Wolfram (1984).

16 Langton (1989/1996).

17 El parámetro lambda oscila entre 0 y 1. Se calcula tomando una célula o estado del autómata celular al azar; hallando (normalmente mediante búsqueda empírica) el número de configuraciones vecinas posibles que “enciende” la célula y dividiendo esto entre el número total posible de configuraciones vecinas. Langton averiguó que surgían estructuras relativamente estables, aunque flexibles cuando lambda se aproximaba a 0,273.

18 Otra medida de la complejidad, el “parámetro Z”, ha sido definida por Andrew Wuensche. Las dos medidas están muy relacionadas, pero la de Wuensche es más general: permite ciertas excepciones de lambda.

19 Linsker (1988).

VI. PERO ¿ES INTELIGENCIA DE VERDAD?

1 Mi discusión resume muchos argumentos complejos e ideas difíciles. Se dan más detalles en Boden (2006), capítulo 16, y, por supuesto, en los escritos de los filósofos mencionados.

2 Turing (1950).

3 Colby et al. (1972).

4 Chalmers (1995, 1996).

5 McGinn (1991).

6 Fodor (1992).

7 Franklin (2007); Franklin et al. (2007).

8 Baars (1988).

9 Otras influencias de IA en LIDA son el sistema Copycat para encontrar analogías creativas; las arquitecturas de subsunción en robótica; el aprendizaje bayesiano multicapa de patrones espaciales; y el marco de CogAff para inteligencia emocional: véanse los capítulos III-V.

10 Churchland (1981, 1986).

11 Dennett (1991).

12 Sloman y Chrisley (2003).

13 Dennett (1995)

14 Putnam (1960).

15 Block (1978).

16 Newell y Simon (1976); Newell (1980).

17 Wittgenstein (1953).

18 Kirsh (1991).

19 Searle (1980).

20 Millikan (1984).

21 Clark (1997); Wheeler (2014).

22 Gallagher (2014).

23 Franklin (1997).

24 Minsky (1965); Boden (2006), 7, i, g.

25 Dennett (1984).

26 Dennett (1991), capítulo 13.

27 Hofstadter (2007).

28 Putnam (1964).

29 Jonas (1966/2001). Y sigue a los cibernéticos al dar un enfoque en lo fundamental similar de la homeostasis fisiológica y el ciclo de acción / percepción (psicológico).

30 Friston (2006, 2013). El principio de energía libre establece que los organismos vivos (y el cerebro) conservan su orden interior

maximizando la utilidad esperada (recompensa) y minimizando el error de predicción (coste). Ofrece en líneas generales una explicación bayesiana sobre cómo permite el cerebro la percepción y el aprendizaje.

31 Boden (1999).

32 Véase Boden (2006) para una discusión sobre este impasse filosófico, 16, pp. vi-viii.

33 Putnam (1997, 1999).

34 Smith (1996).

VII. LA SINGULARIDAD

1 Russell y Norvig (2013).

2 Good (1965).

3 Vinge (1983, 1993).

4 Kurzweil (2005, 2008).

5 Bostrom (2014).

6 Clark (2003, 2008).

7 Hansell y Grassie (2011).

8 Thompson (1984).

9 Brynjolfsson y McAfee (2014).

10 Wilks (2010).

11 Los comentarios de los medios sobre este asunto son legión: intente el lector googlear "sexo con robots". Muchos de estos reaccionan a un libro de un investigador de IA importante (Levy, 2007). Para un análisis filosófico excelente del concepto del amor personal que demuestra cómo de radicalmente difiere de la lujuria y la obsesión sexual, véase Fisher (1990). Aunque este análisis no está escrito desde un punto de vista informático, concuerda con el modelo de mente / identidad favorecido por la IA: véase el capítulo VI.

12 Weizenbaum (1976).

13 Boden (1977/1987). Como ha señalado un revisor, esto, en un sentido, es de lo que trata el libro. Por el contrario (Winston, 1977) publicó en el mismo mes, concentrado en cómo enseñar a hacer IA, no a evaluarla.

14 Sharkey (2012); Sharkey y Sharkey (2012a, b). Estos son solo unos cuantos artículos sobre las implicaciones sociales de la robótica y de la IA en general.

15 Wallach y Allen (2008); Wallach (2015).

16 Whitby (1988, 1996, 2008, 2011).

17 Stuart Russell, comunicación personal.

18 Yudkowsky (2008).

REFERENCIAS

ANDERSON, J. R., The Architecture of Cognition, Cambridge, MA, Harvard University Press, 1983.

—, y lebiere, C., “The Newell Test for a Theory of Cognition”, Behavioral and Brain Sciences, 26, 2003, pp. 587-640.

ASHBY, W. R., “The Nervous System as a Physical Machine: with Special Reference to the Origin of Adaptive Behaviour”, Mind, 56, 1947, pp. 44-59.

—, Design for a Brain: The Origin of Adaptive Behaviour, Londres, Wiley, 1952; 2ª ed., rev., Londres, Chapman, 1960.

BAARS, B. J., A Cognitive Theory of Conscience, Cambridge, Cambridge University Press, 1988.

BAHDANAU, D., cho, K. y BENGIO, Y., “Neural Machine Translation by Jointly Learning to Align and Translate”, International Conference on Learning Representations, http://arxiv.org/abs/1409.0473, 2015.

BAKER, S., Final Jeopardy: The Story of Watson, the Computer That Will Transform Our World, Boston, MA, Mariner Books, 2012.

BARTLETT, J., REFFIN, J., RUMBALL, N. y WILLIAMSON, S., Anti-Social Media, Londres, DEMOS, 2014.

BEER, R. D., Intelligence as Adaptive Behavior: An Experiment in Computational Neuroethology, Boston, MA, Academic Press, 1990.

BENGIO, Y., DUCHARME, R., VINCENT, P. y JAUVIN, C., "A Neural Probabilistic Language Model", Journal of Machine Learning Research, 3, 2003, pp. 1137-1155.

BEURLE, R. L., "Properties of a Mass of Cells Capable of Regenerating Pulses", Philosophical Transactions of the Royal Society, B, 240, 1956, pp. 55-59.

BIRD, J. y layzell, P., "The Evolved Radio and its Implications for Modelling the Evolution of Novel Sensors", Proceedings of Congress on Evolutionary Computation, CEC-2002, pp. 1836-1841.

BLAKE, D. V. y UTTLEY, A. M., eds., The Mechanization of Thought Processes, 2 vols., Londres, Her Majesty's Stationery Office, 1959.

BLOCK, N., "Troubles with Functionalism", en SAVAGE, C. W., ed., Perception and Cognition: Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science, 9, Minneapolis, University of Minnesota Press, 1978, pp. 261-325.

BODEN, M. A., Artificial Intelligence and Natural Man, Nueva York, Basic Books, 1977. [Inteligencia artificial y hombre natural, Julio C. Armero Sanjosé, tr., Madrid, Tecnos, 1983].

—, The Creative Mind: Myths and Mechanisms, 1990/2004; 2ª ed., Londres, Routledge, 2004. [La mente creativa, mitos y mecanismos, José Ángel Álvarez, tr., Barcelona, Gedisa, 2009].

—, (ed.), The Philosophy of Artificial Life, Oxford, Oxford University Press, 1996. [Filosofía de la inteligencia artificial, México DF, FCE, 1996].

—, "Is Metabolism Necessary?", British Journal for the Philosophy of Science, 50/2, 1999, pp. 231-248; reimpresión, BODEN, M. A., Creativity and Art: Three Roads to Surprise, Oxford, Oxford University Press, 2010, pp. 235-254.

—, Mind as Machine: A History of Cognitive Science, 2 vols., Oxford, Oxford University Press, 2002.

—, y EDMONDS, E. A., “What Is Generative Art?”, Digital Creativity, 20(1-2), 2009, pp. 21-46; reimpresión, M. A. BODEN, Creativity and Art: Three Roads to Surprise, Oxford, Oxford University Press, 2010, pp. 125-163.

BOSTROM, N., Superintelligence: Paths, Dangers, Strategies, Oxford, Oxford, University Press, 2014.

BOTTOU, L., “From Machine Learning to Machine Reasoning”, Machine Learning, 94, 2014, pp. 133-149.

BOUKHTOUTA, A., BERGER, J., GUITOUNI, A., BOUAK, F. y BEDROUNI, A., Description and Analysis of Military Planning Systems, Quebec, Canadian Defence Research and Development Technical Report, 2005.

BROOKS, R. A., “Intelligence without Representation”, Artificial Intelligence, 47, 1991, pp. 139-159.

BRYNJOLFSSON, E. y MCAFEE, A., 2014, The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies, Nueva York, W. W. Norton, 2014.

CHALMERS, D. J., “Facing Up To the Problem of Conscience”, Journal of Conscience Studies, 2, 1995, pp. 200-219.

—, The Conscious Mind: In Search of a Fundamental Theory, Oxford, Oxford University Press, 1996.

CHURCHLAND, P. M., “Eliminative Materialism and the Propositional Attitudes”, Journal of Philosophy, 78, 1981, pp. 67-90.

—, "Cognitive Neurobiology: A Computational Hypothesis for Laminar Cortex", Biology and Philosophy, 1, 1986, pp. 25-51.

—, A Neurocomputational Perspective: The Nature of Mind and the Structure of Science, Cambridge, MA, MIT Press, 1989.

CLARK, A. J., Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing, Cambridge, MA, MIT Press, 1989.

—, Associative Engines: Connectionism, Concepts and Representational Change, Cambridge, MA, MIT Press, 1993.

—, Being There: Putting Brain, Body and World Together Again, Londres, MIT Press, 1997.

—, Natural-Born Cyborgs: Why Minds and Technologies Are Made to Merge, Oxford, Oxford University Press, 2003.

—, Supersizing the Mind: Embodiment, Action and Cognitive Extension, Oxford, Oxford University Press, 2008.

—, y KARMILOFF-SMITH, A., "The Cognizer's Innards: A Psychological and Philosophical Perspective on the Development of Thought", Mind and Language, 8, 1993, pp. 487-519.

—, y Thornton, C., "Trading Spaces: Computation, Representation, and the Limits of Uninformed Learning", Behavioral and Brain Sciences, 20, 1997, pp. 57-90.

CLIFF, D., HARVEY, I. y HUSBANDS, P., "Explorations in Evolutionary Robotics", Adaptive Behavior, 2, 1993, pp. 73-110.

COHEN, H., "The Further Exploits of aaron Painter", en FRANCHI, S. y GUZELDERE, G., eds., Constructions of the Mind: Artificial Intelligence and the Humanities, Stanford University Review, número especial, 4/2, 1995, pp. 1-345, y 141-160.

COHEN, H., “A Million Millennial Medicis”, en CANDY, L. y EDMONDS, E., eds., Explorations in Art and Technology, Londres, Springer, 2002, pp. 91-104.

COLBY, K. M., “Computer Simulation of a Neurotic Process”, en TOMKINS, S. S. y MESSICK, S., eds., Computer Simulation of Personality: Frontier of Psychological Research, Nueva York, Wiley, 1963, pp. 165-180.

—, HILF, F. D., WEBER, S. y KRAMER, H. C., “Turing-Like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes”, Artificial Intelligence, 3, 1972, pp. 199-222.

—, WEBER, S. y HILF, F. D., “Artificial Paranoia”, Artificial Intelligence, 2, 1971, pp. 1-6.

COLLOBERT, R., WESTON, J., BOTTOU, L., KARLEN, M., KAVUKCUOGLU, K. y KUKSU, P., “Natural Language Processing (Almost) from Scratch”, Journal of Machine Learning Research, 12, 2011, pp. 2493-2537.

COLTON, S., “The Painting Fool: Stories from Building an Automated Painter”, en MCCORMACK, J. y D’INVERNO, M., eds., Computers and Creativity, Londres, Springer, 2012, pp. 3-38.

COOPER, R., FOX, J., FARRINGDON, J. y SHALLICE, T., “Towards a Systematic Methodology for Cognitive Modelling”, Artificial Intelligence, 85, 1996, pp. 3-44.

—, SCHWARTZ, M., YULE, P. y SHALLICE, T., “The Simulation of Action Disorganization in Complex Activities of Daily Living”, Cognitive Neuropsychology, 22, 2005, pp. 959-1004.

COPE, D., Computer Models of Musical Creativity, Cambridge, MA, MIT Press, 2005.

CRAIK, K. J. W., The Nature of Explanation, Cambridge, Cambridge University Press, 1943.

DAVEY, A. C., Discourse Production: A Computer Model of Some Aspects of a Speaker, Edimburgo, Edinburgh University Press, 1978.

DAYAN, P. y ABBOTT, L., Theoretical Neuroscience: Computational and Mathematical Modelling of Neural Systems, Cambridge, MA, MIT Press, 2001.

DEATON, C., LENAT, D. et al. (8 autores), “The Comprehensive Terrorism Knowledge Base in Cyc”, Proceedings of the 2005 International Conference on Intelligence Analysis, McLean, VA, mayo de 2005.

DENNETT, D. C., Elbow Room: The Varieties of Free Will Worth Wanting, Cambridge, MA, MIT Press, 1984.

—, Consciousness Explained, Londres, Allen Lane, 1991.

—, “The Unimagined Preposterousness of Zombies”, Journal of Conscience Studies, 2, 1995, pp. 322-326.

FISHER, M., Personal Love, Londres, Duckworth, 1990.

FODOR, J. A., “The Big Idea: Can There Be a Science of Mind?”, The Times Literary Supplement, 3 de julio, 1992, pp. 5-7.

FRANKLIN, S., “Autonomous Agents as Embodied AI”, Cybernetics and Systems, Special issue on Epistemological Aspects of Embodied AI, 28(6), 1997, pp. 499-520.

—, “A Foundational Architecture for Artificial General Intelligence”, en GOERTZEL, B. y WANG, P., eds., Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms, Amsterdam, IOS Press, 2007, pp. 36-54.

—, et al. (7 autores), "LIDA: A Computational Model of Global Workspace Theory and Developmental Learning", AAIA Fall Symposium on IA and Conscience: Theoretical Foundations and Current Approaches, Arlington, VA, AAAI, 2007.

FRISTON, K., "A Free-Energy Principle for the Brain?", Journal of Physiology, París, 100, 2006, pp. 70-87.

—, "Life As We Know It", J. R. Soc. Interface, 3 de julio, 10(86), 20130475, 2013.

GALLAGHER, S., "Phenomenology and Embodied Cognition", en L. Shapiro, ed., The Routledge Handbook of Embodied Cognition, Londres, Routledge, 2014, pp. 9-18.

GARCIA, C. y DELAKIS, M., "Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, 26, 2004, pp. 1408-1423.

GARDNER, M., "The Fantastic Combinations of John Conway's New Solitaire Game 'Life'", Scientific American, 223/4, 1970, pp. 120-123.

GIBSON, J. J., The Ecological Approach to Visual Perception, Londres, Houghton Mifflin, 1979.

GIGERENZER, G., "Fast and Frugal Heuristics: The Tools of Bounded Rationality", en KOEHLER, D. J. y HARVEY, N., eds., Blackwell Handbook of Judgment and Decision Making, Oxford, Blackwell, 2004, pp. 62-88.

GOOD, I. J., "Speculations Concerning the First Ultraintelligent Machine", en ALT, F. y RUBINOFF, M., eds., Advances in Computing, vol. 6, Nueva York, Academic, 1965, pp. 30-38.

GOODWIN, B. C., How the Leopard Changed Its Spots: The Evolution of Complexity, princeton, nj, Princeton University Press, 1994. [Las manchas del leopardo. La evolución de la complejidad, Ambrosio García, tr., Barcelona, Tusquets, 2008].

GRAVES, A., MOHAMED, A.-R. y HINTON, G. E., "Speech Recognition with Deep Recurrent Neural Networks", Proc. International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645-6649.

GREY WALTER, W., "An Imitation of Life", Scientific American, 182 (5), 1950, pp. 42-45.

HANSELL, G. R. y GRASSIE, W., eds., H+/-: Transhumanism and its Critics, Filadelfia, PA, Metanexus, 2011.

HARVEY, I., HUSBANDS, P. y CLIFF, D., "Seeing the Light: Artificial Evolution, Real Vision", en CLIFF, D., HUSBANDS, P., MEYER, J.-A. y WILSON, S. W., eds., From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior, Cambridge, MA, MIT Press, 1994, pp. 392-401.

HAUGELAND, J., "The Nature and Plausibility of Cognitivism", Behavioral and Brain Sciences, 1, 1978, pp. 215-226.

HINTON, G. E., OSINDERO, S. y TEH, Y.-W., "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, 18, 2006, pp. 1527-1554.

HINTON, G. E. y SEJNOWSKI, T. J., "Learning and Relearning in Boltzmann Machines", 1985; en RUMELHART, D. E. y MCCLELLAND, J. L., eds., Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations,Cambridge, MA, MIT Press, 1986a, pp. 282-317.

HINTON, G. E., DENG, L., yu, D., et al. (11 autores), "Deep Neural Networks for Acoustic Modeling in Speech Recognition",

Signal Processing Magazine, IEEE, 29(6), 2012, pp. 82-97.

HOFSTADTER, D. R., I Am a Strange Loop, Nueva York, Basic Books, 2007.

—, y Mitchell, M., “The Copycat Project: A Model of Mental Fluidity and Analogy-Making”, en HOFSTADTER, D. y FARG, Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought, Nueva York, Basic Books, 1997, pp. 205-300.

HSU, F., Behind Deep Blue: Building the Computer that Defeated the World Chess Champion, Princeton, NJ, Princeton University Press, 2002.

HUTCHINS, E. L., Cognition in the Wild, Cambridge, MA, MIT Press, 1995.

JONAS, H., The Phenomenon of Life: Toward a Philosophical Biology, Nueva York: Harper Collins, 1966/2001; reimpresión, Evanston, IL, Northwestern University Press, 2001.

KIRSH, D., “Today the Earwig, Tomorrow Man?”, Artificial Intelligence, 47, 1991, pp. 161-184.

KOLODNER, J. L., “An Introduction to Case-Based Reasoning”, Artificial Intelligence Review, 6, 1992, pp. 3-34.

KONING, H. y EIZENBERG, J., “The Language of the Prairie: Frank Lloyd Wright’s Prairie Houses”, Environment and Planning, B, 8, 1981, 295-323.

KRISHEVSKY, A., SUTSKEVER, I., y HINTON, G. E., “ImageNet Classification with Deep Convolutional Neural Networks”, Advances in Neural Information Processing Systems, 25, 2012, pp. 1090-1098.

KURZWEIL, R., The Singularity is Near: When Humans Transcend Biology, Londres, Penguin, 2005. [La Singularidad está cerca. Cuando los humanos trascendamos la biología, Carlos García Hernández, tr., Berlín, Lola Books, 2012].

—, The Age of Spiritual Machines: When Computers Exceed Human Intelligence, Londres, Penguin, 2008.

LAIRD, J. E., NEWELL, A. y ROSENBLOOM, P., “Soar: An Architecture for General Intelligence”, Artificial Intelligence, 33, 1987, pp. 1-64.

LANGTON, C. G., “Artificial Life”, en LANGTON, C. G. (ed.), Artificial Life, Redwood City, ca, Addison-Wesley, 1989, pp. 1-47; versión revisada en BODEN, M. A., ed., The Philosophy of Artificial Life, Oxford, Oxford University Press, 1996, pp. 39-94. [Filosofía de la inteligencia artificial, México DF, FCE, 1996].

—, “Life at the Edge of Chaos”, en LANGTON, C. J., TAYLOR, C., FARMER, J. D, y RASMUSSEN, S., eds., Artificial Life II, Redwood City, CA, Addison-Wesley, 1991, pp. 41-91.

LECUN, Y., BENGIO, Y. y HINTON, G. E., “Deep Learning”, Nature, 521, 2015, pp. 436-444.

LENAT, D. B., “cyc: A Large-Scale Investment in Knowledge Infrastructure”, en Communications of the Association for Computing Machinery, 38(11), pp. 1995, 33-38.

LEVY, D., Love and Sex with Robots: The Evolution of Human-Robot Relationships, Londres, Duckworth, 2007.

LINSKER, R., “Self-Organization in a Perceptual Network”, Computer, 21, 1988, pp. 105-117; reimpresión, ANDERSON, J. A., PELLIONISZ, A. y ROSENFELD, E., eds., Neurocomputing 2: Directions for Research, Cambridge, MA, MIT Press, pp. 528-540.

LOVELACE, A. A., “Notes by the Translator”, 1843; reimpresión, HYMAN, R. A., (ed.), Science and Reform: Selected Works of Charles Babbage, Cambridge, Cambridge University Press, 1989, pp. 267-311.

MACKAY, D. M., “On the Combination of Digital and Analogue Computing Techniques in the Design of Analytic Engines”, en BLAKE, D. V. y UTTLEY, A. M., eds., The Mechanization of Thought Processes, vol. 1, Londres, Her Majesty’s Stationery Office, 1949/1959, pp. 55-65; apareció por primera vez para consumo privado en mayo de 1949.

MCCARTHY, J., “Programs with Common Sense”, en BLAKE, D. V. y UTTLEY, A. M., eds., The Mechanization of Thought Processes, vol. 1, Londres, Her Majesty’s Stationery Office, 1959, pp. 75-91.

—, Situations, Actions and Causal Laws, Stanford, ca, Stanford IA Project Memo 2, 1963; reimpreso como la sección 7.2 de MCCARTHY, J., “Programs with Common Sense”, en MINSKY, M. L., (ed.), Semantic Information Processing, Cambridge, MA, MIT Press, 1968, pp. 403-417.

—, “Circumscription–A Form of Non-Monotonic Reasoning”, Artificial Intelligence, 13, 1980, pp. 27-24.

—, “Applications of Circumscription to Formalizing Common-Sense Knowledge”, Artificial Intelligence, 28, 1986, pp. 86-116.

—, y HAYES, P. J., “Some Philosophical Problems from the Standpoint of Artificial Intelligence”, en MELTZER, B. y MICHIE, D. M., eds., Machine Intelligence 4, Edimburgo, Edinburgh University Press, 1969, pp. 463-502.

MCCLELLAND, J. L. y RUMELHART, D. M. (y el Grupo de Investigación PDP), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 2, Psychological and Biological Models, Cambridge, MA, MIT Press, 1986.

MCCORMACK, J., Impossible Nature: The Art of Jon McCormack, Melbourne, Australian Centre for the Moving Image, 2004.

MCCULLOCH, W. S. y PITTS, W. H., “A Logical Calculus of the Ideas Immanent in Nervous Activity”, Bulletin of Mathematical Biophysics, 5, 1943, pp. 115-133; reimpresión, papert, S., ed., Embodiments of Mind, Cambridge, MA, MIT Press, 1965, pp. 19-39.

MCGINN, C., The Problem of Conscience, Oxford, Basil Blackwell, 1991.

MARR, D. C., Vision: A Computational Investigation into the Human Representation and Processing of Visual Información, San Francisco, CA, Freeman, 1982.

MICHALSKI, R. S. y CHILAUSKY, R. L., “Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis”, International Journal of Policy Analysis and Information Systems, 4, 1980, pp. 125-161.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. y DEAN, J., “Distributed Representations of Words and Phrases and their Compositionality”, Advances in Information Processing Systems, 26, 2013; 3111-3119.

MILLIKAN, R. G., Language, Thought and Other Biological Categories: New Foundations for Realism, Cambridge, MA, MIT Press, 1984.

MINSKY, M. L., “Steps Toward Artificial Intelligence [publicado originalmente en 1956 como un informe técnico en el MIT: Heuristic Aspects of the Artificial Intelligence Problem]”, Proceedings of the Institute of Radio Engineers, 49, 1956/1961, pp. 8-30; reimpresión,

E. A. FEIGENBAUM y J. A. FELDMAN, eds., Computers and Thought, Nueva York, McGraw-Hill, 1963, pp. 406-450.

—, “Matter, Mind and Models”, Proceedings of the International Federation of Information Processing Congress, 1, Washington DC, Spartan, 1965, pp. 45-49.

—, The Society of Mind, Nueva York, Simon & Schuster, 1985.

—, The Emotion Machine: Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind, Nueva York, Simon and Schuster, 2006. [La máquina de las emociones. Sentido común, inteligencia artificial y el futuro de la mente humana, Mercedes García Garmilla, tr., Barcelona, Debate, 2010].

—, y PAPERT, S. A., Perceptrons: An Introduction to Computational Geometry, Cambridge, MA, MIT Press, 1969.

—, y PAPERT, S. A., “Prologue: A View From 1988” y “Epilogue: The New Connectionism”, en Perceptrons: An Introduction to Computational Geometry, 2ª ed., Cambridge, MA, MIT Press, VIII-XV y 247-280, 1988.

MITCHELL, T. M, Machine Learning, Nueva York, McGraw-Hill, 1997.

—, “The Discipline of Machine Learning”, Technical Report CMU-ML-06-108, Pittsburgh, PA, Carnegie Mellon University School of Computer Science, 2006.

MNIH, V., HASSABIS, D. et al. (19 autores), “Human-Level Control Through Deep Reinforcement Learning”, Nature, 518, 2015, pp. 529-533.

NEISSER, U., “The Imitation of Man by Machine”, Science, 139, 1963, pp. 193-197.

NEWELL, A., “Physical Symbol Systems”, Cognitive Science, 4, 1980, pp. 135-183.

—, Unified Theories of Cognition, Cambridge, MA, Harvard University Press, 1990.

—, SHAW, J. C. y SIMON, H. A., “A General Problem-Solving Program for a Computer”, Proceedings of the International Conference on Information Processing, París, junio, 1959, pp. 256-264.

NEWELL, A. y SIMON, H. A., “The Logic Theory Machine”, IRE Transactions on Información Theory, it-2(3), 1956, pp. 61-79.

—, Human Problem Solving, Englewood Cliffs, NJ, Prentice-Hall, 1972.

—, “Computer Science as Empirical Enquiry: Symbols and Search”, Communications of the Association for Computing Machinery, 19, 1976, pp. 113-126.

PHILIPPIDES, A., HUSBANDS, P. y O’SHEA, M., “Neural Signalling–It’s a Gas!”, en NIKLASSON, L., BODEN, M. y ZIEMKE, T., eds., ICANN98: Proceedings of the 8th International Conference on Artificial Neural Networks, Londres, Springer-Verlag, 1998, pp. 51-63.

PHILIPPIDES, A., HUSBANDS, P., SMITH, T. y o’SHEA, M., “Flexible Couplings: Diffusing Neuromodulators and Adaptive Robotics”, Artificial Life, 11, 2005, pp. 139-160.

PICARD, R. W., Affective Computing, Cambridge, MA, MIT Press, 1997.

—, “Response to Sloman’s Review of Affective Computing”, AI Magazine, 20(1) (marzo), 1999, pp. 134-137.

PINKER, S. y PRINCE, A., “On Language and Connectionism: Analysis of a Parallel Distributed Model of Language Acquisition”, Cognition, 28, 1998, pp. 73-193.

PITTS, W. H. y MCCULLOCH, W. S., “How We Know Universals: The Perception of Auditory and Visual Forms”, Bulletin of Mathematical Biophysics, 9, 1947, pp. 127-147; reimpresión, S. Papert, ed., Embodiments of Mind, Cambridge, MA, MIT Press, 1965, pp. 46-66.

PUTNAM, H., “Minds and Machines”, en Hook, S., ed., Dimensions of Mind: A Symposium, Nueva York, New York University Press, 1966, pp. 148-179.

—, “Robots: Machines or Artificially Created Life?”, The Journal of Philosophy, 61, 1964, pp. 668-691.

—, “Functionalism: Cognitive Science or Science Fiction?”, en JOHNSON, D. M. y ERNELING, C. E., eds., The Future of the Cognitive Revolution, Oxford, Oxford University Press, 1997, pp. 32-44.

—, The Threefold Cord: Mind, Body and World, Nueva York, Columbia University Press, 1999.

RAY, T. S., “An Approach to the Synthesis of Life”, en LANGTON, C. J., TAYLOR, C., FARMER, J. D. y RASMUSSEN; S., eds., Artificial Life II, Redwood City, CA, Addison-Wesley, 1991, pp. 371-408.

ROSENBLATT, F., “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”, Psychological Review, 65, 1958, pp. 386-408.

—, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Washington DC, Spartan, 1962.

RUMELHART, D. E. y MCCLELLAND, J. L. (y el grupo de investigación PDP), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, Foundations, Cambridge, MA, MIT Press, 1986a.

—, y MCCLELLAND, J. L., “On Learning the Past Tense of English Verbs”, en RUMELHART, D. E. y MCCLELLAND, J. L., eds., Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, Foundations, Cambridge, MA, MIT Press, 1986b, pp. 216-271.

RUSSELL, S. y NORVIG, P., Artificial Intelligence: A Modern Approach, 3ª ed., Londres, Pearson, 2013. [Inteligencia artificial: un enfoque moderno, Juan Manuel Corbacho Rodríguez, tr., Madrid, Alhambra, 2004].

SAHOTA, M. y MACKWORTH, A. K., “Can Situated Robots Play Soccer?”, Proceedings of the Canadian Conference on Artificial Intelligence, Banff, Alberta, 1994, pp. 249-254.

SAMPSON, G. R., The ‘Language Instinct’ Debate: Revised Edition, Londres, Continuum, 2005.

SAMUEL, A. L., “Some Studies in Machine Learning Using the Game of Checkers”, IBM Journal of Research and Development, 3, 1959, pp. 211-229; reimpresión, E. A. FEIGENBAUM y FELDMAN, J. A., eds., Computers and Thought, Nueva York, McGraw-Hill, 1963, pp. 71-108.

SCHANK, R. C. y ABELSON, R. P., Scripts, Plans, Goals and Understanding, Hillsdale, NJ, Lawrence Erlbaum, 1977. [Guiones, planes, metas y entendimiento, Elisabeth Gilboy y Javier Zanón, trs., Barcelona, Paidós, 1988].

SEARLE, J. R., “Minds, Brains and Programs”, Behavioral and Brain Sciences, 3, 1980, pp. 417-457, 1980. Incluye comentarios de colegas y respuestas.

SELFRIDGE, O. G., “Pandemonium: A Paradigm for Learning”, en BLAKE, D. V. y UTTLEY, A. M., eds., The Mechanization of Thought Processes, vol. 1, Londres, Her Majesty’s Stationery Office, 1959, pp. 511-529.

SHALLICE, T. y COOPER, R. P., The Organisation of Mind, Oxford, Oxford University Press, 2011.

SHANAHAN, M., Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia, Cambridge, MA, MIT Press, 1997.

SHARKEY, N. E., “Killing Made Easy: From Joystics to Politics”, en SCHEIPERS, S. y STRECHAN, H., eds., The Changing Character of War, Oxford, Oxford University Press, 2012.

—, y SHARKEY, A. J. C., “Robot Surgery and Ethical Challenges”, en GOMES, P. (ed.), Medical Robotics: Minimally Invasive Surgery, Salt Lake City, UT, Woodland Publishing, 2012a, pp. 276-291.

—, y SHARKEY, A. J. C., “The Rights and Wrongs of Robot Care”, en LIN, P., ABNEY, K. y BEKEY, G. A., eds., Robot Ethics: The Ethical and Social Implications of Robotics, Cambridge, MA, MIT Press, 2012b, pp. 267-282.

SIMON, H. A., “Motivational and Emotional Controls of Cognition”, Psychological Review, 74, 1967, pp. 39-79.

SLOMAN, A., The Computer Revolution in Philosophy (y otros muchos artículos acerca de la IA), publicado originalmente en 1978, Brighton, Harvester, pero actualizado constantemente: http://www.cs.bham.ac.uk/research/cogaff/crp/, sin fecha.

—, “Interactions Between Philosophy and Artificial Intelligence: The Role of Intuition and Non-Logical Reasoning in Intelligence”,

Artificial Intelligence, 2, 1971, pp. 209-225.

—, “Afterthoughts on Analogical Representation”, en SCHANK, R. C. y NASH-WEBBER, B. L., eds., Theoretical Issues in Natural Language Processing, Arlington, va, Association for Computational Linguistics, 1975, pp. 164-168.

—, “Review of R. Picard, Affective Computing”, IA Magazine, 20/1, 1999, pp. 127-133.

—, y CHRISLEY, R. L., “Virtual Machines and Conscience”, en HOLLAND, O., ed., Machine Conscience, Exeter, Imprint Academic, Journal of Conscience Studies, número especial, 10(4/5), 2003, pp. 133-172.

SMITH, B. C., On the Origin of Objects, Cambridge, MA, MIT Press, 1996.

SUN, R., SLUSARZ, P. y TERRY, C., “The Interaction of the Explicit and the Implicit in Skill Learning: A Dual-Process Approach”, Psychological Review, 112, 2005, pp. 159-192.

SUTSKEVER, I., MARTENS, J. y HINTON, G. E., “Generating Text with Recurrent Neural Networks”, Proc. 28th International Conference on Machine Learning, 2011, pp. 1017-1124.

SUTSKEVER, I., VINTALS, O. y LE, Q. V., “Sequence to Sequence Learning with Neural Networks”, Advances in Neural Information Processing Systems, 27, 2014, pp. 3104-3012.

TAIGMAN, Y., YANG, M., ranzato, M. y WOLF, L., “Deepface: Closing the Gap to Human-Level Performance in Face Verification”, Proc. Conference on Computer Vision and Pattern Recognition, http://arxiv.org/abs/1411.4280, 2014 (visitado el 2 de diciembre de 2015).

TAUBE, M., Computers and Common Sense: The Myth of Thinking Machines, Nueva York, Columbia University Press, 1961.

THOMPSON, H., “There Will Always Be Another Moonrise: Computer Reliability and Nuclear Weapons”, The Scotsman, 17 de octubre de 1984, p. 11; reimpresión: AISB Quarterly, 53-54 (Spring-Summer), pp. 21-23, disponible en http://www.aisb.org.uk/articles/moonrise.html.

THORNTON, C., Truth from Trash: How Learning Makes Sense, Cambridge, MA, MIT Press, 2000.

TODD, S. C. y LATHAM, W., Evolutionary Art and Computers, Londres, Academic Press, 1992.

TURING, A. M., “On Computable Numbers with an Application to the Entscheidungsproblem”, Proceedings of the London Mathematical Society, Serie 2, 42/3 y 42/4, 1936; reimpresión, DAVIS, M., ed., The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions, Hewlett, NY, Raven Press, 1965, pp. 116-153.

—, “Computing Machinery and Intelligence”, Mind, 59, 1950, pp. 433-460; reimpresión, BODEN, M. A., ed., The Philosophy of Artificial Intelligence, Oxford, Oxford University Press, 1990, pp. 40-66. [Filosofía de la inteligencia artificial, México, DF, FCE, 1996].

—, “The Chemical Basis of Morphogenesis”, en Philosophical Transactions of the Royal Society: B, 237, 1952, pp. 37-72.

TURK, G., “Generating Textures on Arbitrary Surfaces Using Reaction-Diffusion”, Computer Graphics, 25, 1991, pp. 289-298.

UTTLEY, A. M., “Conditional Probability Machines and Conditioned Reflexes”, en C. E. SHANNON y J. MCCARTHY, eds., Automata Studies, Princeton, NJ, Princeton University Press, 1956, pp. 253-275.

—, “Conditional Probability Computing in a Nervous System”, en BLAKE, D. V. y UTTLEY, A. M., eds., The Mechanization of Thought Processes, vol. 1, Londres, Her Majesty’s Stationery Office, 1959, pp. 119-147.

VINCZE, M., WACHSMUTH, S. y SAGERER, G., “Perception and Computer Vision”, en FRANKISH, K. y RAMSEY, W. M., eds., The Cambridge Handbook of Artificial Intelligence Cambridge, Cambridge University Press, 2014168-90.

VINGE, V., “First Word”, OMNI Magazine, enero, 10, 1983.

—, “The Coming Technological Singularity”, Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace, NASA Conference Publication 10129, 1993, pp. 11-22.

VON NEUMANN, J., “The General and Logical Theory of Automata”, en JEFFRESS, L. A. (ed.), Cerebral Mechanisms in Behavior: The Hixon Symposium, Nueva York, Wiley, 1951, pp. 1-13.

WALLACH, W., A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control, Oxford, Oxford University Press, 2015.

—, y allen, C., Moral Machines: Teaching Robots Right From Wrong, Oxford, Oxford University Press, 2008.

WEBB, B., “A Cricket Robot”, Scientific American, 275(6), 1996, pp. 94-99.

WEIZENBAUM, J., Computer Power and Human Reason: From Judgment to Calculation, San Francisco, W. H. Freeman, 1976.

WHEELER, M., “Revolution, Reform, or Business as Usual? The Future Prospects for Embodied Cognition”, en SHAPIRO, L. (ed.),

The Routledge Handbook of Embodied Cognition, Londres: Routledge, 2014, 374-383.

WHITBY, B., Artificial Intelligence: A Handbook of Professionalism, Chichester, Ellis Horwood, 1988.

—, Reflections on Artificial Intelligence: The Social, Legal and Moral Dimensions, Oxford, Intellect Books, 1996.

—, “Sometimes It’s Hard to be a Robot: A Call for Action on the Ethics of Abusing Artificial Agents”, Interacting with Computers, 20(3), 2008, pp. 326-333.

—, “Do You Want a Robot Lover?”, en LIN, P., ABNEY, K. y BEKEY, G. A., eds., Robot Ethics: The Ethical and Social Implications of Robotics, Cambridge, MA, MIT Press, 2011, pp. 233-249.

WHITELAW, M., Metacreation: Art and Artificial Life, Londres, MIT Press, 2004.

WIENER, N., Cybernetics: or Control and Communication in the Animal and the Machine, Cambridge, MA, MIT Press, 1948.

WILKS, Y. A., (ed.), Language, Cohesion and Form: Margaret Masterman 1910-1986, Cambridge, Cambridge University Press, 2005.

—, (ed.), Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues, Amsterdam, John Benjamins, 2010.

WINOGRAD, T., Understanding Natural Language, Edimburgo, Edinburgh University Press, 1972.

WINSTON, P. H., Artificial Intelligence, Reading, MA, Addison-Wesley, 1977.

WITTGENSTEIN, L., Philosophical Investigations, trad. G. E. M. Anscombe, Oxford, Blackwell, 1953. [Investigaciones filosóficas, Jesús Padilla Gálvez, tr., Madrid, Trotta, 2017].

WOLFRAM, S., “Cellular Automata as Models of Complexity”, Nature, 311, 1984, pp. 419-424.

XU, K., BA, J. L., KIROS, R., CHO, K. COURVILLE, A., SALATHUTDINOV, R., ZEMEL, R. S., y BENGIO, Y., (2015), “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, Proc. 32nd International Conference on Machine Learning, http://arxiv.org/abs/1502.03044, 2015, visitado el 2 de diciembre de 2015.

YUDKOWSKY, E., “Artificial Intelligence as a Positive and Negative Factor in Global Risk”, en BOSTROM, N. y CIRKOVIC, M. M., eds., Global Catastrophic Risks, Oxford, Oxford University Press, 2008, pp. 308-345.

¡Gracias por comprar este ebook!

Para ver otros títulos de la colección,
pincha aquí

Síguenos en
www.turnerlibros.com



TurnerLibros



EditorialTurner



editorialturner

